

Manuscript Number: MEEGID-D-18-00328R1

Title: Time-scaled phylogeography of complete Zika virus genomes using discrete and continuous space diffusion models

Article Type: Research paper

Keywords: Zika virus
Continuous phylogeography
Surveillance
Phylodynamics

Corresponding Author: Professor gianguglielmo zehender, PhD

Corresponding Author's Institution: University of Milan

First Author: Erika Ebranati

Order of Authors: Erika Ebranati; Carla Veo; Valentina Carta; Elena Percivalle; Francesca Rovida; Elena R Frati; Antonella Amendola; Massimo Ciccozzi; Elisabetta Tanzi; Massimo Galli; Fausto Baldanti; gianguglielmo zehender, PhD

Abstract: Zika virus (ZIKV), a vector-borne infectious agent that has recently been associated with neurological diseases and congenital microcephaly, was first reported in the Western hemisphere in early 2015. A number of authors have reconstructed its epidemiological history using advanced phylogenetic approaches, and the majority of Zika phylogeography studies have used discrete diffusion models. Continuous space diffusion models make it possible to infer the possible origin of the virus in real space by reconstructing its ancestral location on the basis of geographical coordinates deduced from the latitude and longitude of the sampling locations. We analysed all the ZIKV complete genome isolates whose sampling times and localities were available in public databases at the time the study began, using a Bayesian approach for discrete and continuous phylogeographic reconstruction.

The discrete phylogeographic analysis suggested that ZIKV emerged to become endemic/epidemic in the first decade of the 1900s in the Ugandan rainforests, and then reached Western Africa and Asia between the 1930s and 1950s. After a long period of about 40 years, it spread to the Pacific islands and reached Brazil from French Polynesia. Continuous phylogeography of the American epidemic showed that the virus entered in north-eastern Brazil in late 2012 and started to spread in early 2013 from two high probability regions: one corresponding to the entire north-east Brazil and the second surrounding the city of Rio de Janeiro, in a mainly northwesterly direction to Central America, the north-western countries of south America and the Caribbean islands. Our data suggest its cryptic circulation in both French Polynesia and Brazil, thus raising questions about the mechanisms underlying its undetected persistence in the absence of a known animal reservoir, and underline the importance of

continuous diffusion models in making more reliable phylogeographic reconstructions of emerging viruses.

Dear Sir,

Please find enclosed a copy of our manuscript entitled “**Time-scaled phylogeography of complete Zika virus (ZIKV) genomes: an investigation of the virus entry into Western hemisphere using discrete and continuous space diffusion models**”, which we would like to submit as a research article for publication in Infection, Genetics and Evolution. The paper has not been previously published or submitted for publication elsewhere.

In this study we showed that Zika virus entered in Brazil in the early 2013, at least two years before the first cases of infection were reported, in a period compatible with an early (still cryptic) diffusion of Zika infection also in French Polynesia. We also made hypothesis about the possible role of alternative routes of transmission of the infection in the cryptic circulation of the virus observed both in Americas and French Polynesia.

Since its first appearance in the Americas, several Authors have attempted to estimate the evolutionary dynamics of the Zika virus by using advanced phylogenetic analysis, including phylogeographical analysis. Nevertheless, the majority of them employed discrete diffusion approaches, requiring the grouping of the isolates on the basis of their sampling localities, which are frequently arbitrarily defined and not precise. In this study we employed a continuous phylogeographical approach which allow the ancestral sequences to reside at any location in a continuous bidimensional space, overcoming the main limit of the discrete approach which is the impossibility to reconstruct the ancestral locality if it is not among the sampled locations.

Our findings show the importance of integrated human, animal and vector surveillance and suggests that phylogenetic studies can contribute to this surveillance.

We believe that our paper will interest other as well as the people involved in public health and prevention strategies against Zika virus.

Yours faithfully,

Gianguglielmo Zehender

Rebuttal letter

Editor

We would like to thank the Academic Editor for his positive consideration in our study.
Changes to the manuscript are listed in the Reply to the Referee

Reply to the Referee

We would like to thank the referee for his positive consideration in our study.

Point-by-point responses :

1. The section 2.6 Recombination detection was moved at point 2.4 and previous 2.4 Likelihood mapping was moved in the place of 2.5.
2. As suggested by the referee, a brief comment to the study of Tower et al has been added to the Discussion section (pag.26 line 208). Four new references, including Tower et al. 2016, have been added (Gao 2016, Olawoyin 2018, Baca-Carrasco 2016).
3. Typos were corrected.

HIGHLIGHTS

- ✓ Discrete and continuous Zika virus phylogeography
- ✓ Importance of continuous phylogeography in the reconstructions of emerging viruses
- ✓ Importance of genome characterisation in the surveillance of emerging infections

1 Time-scaled phylogeography of complete *Zika virus* genomes using discrete and continuous
2 space diffusion models

3

4 Erika Ebranati^{a,b,1}, Carla Veo^{a,b,1}, Valentina Carta^a, Elena Percivalle^c, Francesca Rovida^c, Elena
 5 Rosanna Frati^{b,d}, Antonella Amendola^{b,d}, Massimo Ciccozzi^e, Elisabetta Tanzi^{b,d}, Massimo Galli^{a,b},
 6 Fausto Baldanti^c, Gianguglielmo Zehender^{a,b*}

7

8 ^a Department of Biomedical and Clinical Sciences "L. Sacco", University of Milan, Milano, Italy.

9 ^b CRC-Coordinated Research Center "EpiSoMI", University of Milan, Milano, Italy.

10 ^c Molecular Virology Unit, Microbiology and Virology Department, Fondazione IRCCS Policlinico
 11 San Matteo, Pavia, Italy.

12 ^d Department of Biomedical Sciences for Health, University of Milan, Milano, Italy.

13 ^e Unit of Medical Statistics and Molecular Epidemiology, University Campus Bio-Medico of Rome,
 14 Italy.

15

16 *Corresponding author at: Department of Biomedical and Clinical Sciences "L. Sacco", University
 17 of Milan, Via G.B. Grassi 74, 20157 Milan, Italy

18 E-mail address: gianguglielmo.zehender@unimi.it (G. Zehender)

19

20 ¹ These authors contributed equally to this work.

21

22 **ABSTRACT**

23 *Zika virus* (ZIKV), a vector-borne infectious agent that has recently been associated with
24 neurological diseases and congenital microcephaly, was first reported in the Western hemisphere in
25 early 2015.

26 A number of authors have reconstructed its epidemiological history using advanced phylogenetic
27 approaches, and the majority of Zika phylogeography studies have used discrete diffusion models.
28 Continuous space diffusion models make it possible to infer the possible origin of the virus in real
29 space by reconstructing its ancestral location on the basis of geographical coordinates deduced from
30 the latitude and longitude of the sampling locations. We analysed all the ZIKV complete genome
31 isolates whose sampling times and localities were available in public databases at the time the study
32 began, using a Bayesian approach for discrete and continuous phylogeographic reconstruction.

33 The discrete phylogeographic analysis suggested that ZIKV emerged to become endemic/epidemic
34 in the first decade of the 1900s in the Ugandan rainforests, and then reached Western Africa and
35 Asia between the 1930s and 1950s. After a long period of about 40 years, it spread to the Pacific
36 islands and reached Brazil from French Polynesia. Continuous phylogeography of the American
37 epidemic showed that the virus entered in north-eastern Brazil in late 2012 and started to spread in
38 early 2013 from two high probability regions: one corresponding to the entire north-east Brazil and
39 the second surrounding the city of Rio de Janeiro, in a mainly northwesterly direction to Central
40 America, the north-western countries of south America and the Caribbean islands. Our data suggest
41 its cryptic circulation in both French Polynesia and Brazil, thus raising questions about the
42 mechanisms underlying its undetected persistence in the absence of a known animal reservoir, and
43 underline the importance of continuous diffusion models in making more reliable phylogeographic
44 reconstructions of emerging viruses.

45

46 **KEYWORDS**

- 47 Zika virus
- 48 Continuous phylogeography
- 49 Surveillance
- 50 Phylodynamics

51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75

1. INTRODUCTION

Zika virus (ZIKV) is an emerging arbovirus belonging to the *Flaviviridae* family, genus *Flavivirus*, and is closely phylogenetically related to other important mosquito-borne flaviviruses such as *Japanese encephalitis*, *West Nile* and *Dengue* viruses (Gubler et al., 2017). It was first discovered in a rhesus macaque monkey with fever kept in captivity in the Zika forest on the Entebbe peninsula in Uganda in 1947 (Buechler et al., 2017).

The viral genome is represented by a single-stranded positive-sense RNA molecule of 10.7 kbp that encodes a single polyprotein encompassing three structural proteins (the capsid, the precursor membrane and the envelope proteins) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5), which play essential roles in virus replication, virulence and secretion (Mittal et al., 2017). Phylogenetic studies of whole ZIKV genomes have revealed the existence of two major evolutionary lineages: one African and the other Asian (Simonin et al., 2017), encompassing also isolates from Pacific Islands and America.

ZIKV is naturally maintained by two distinct transmission cycles: a sylvatic cycle involving non-human primates and arboreal mosquitoes of the genus *Aedes*, and an urban cycle involving humans and urban mosquitoes (mainly *A. aegypti*). It can also be transmitted without vectors: vertically from an infected mother to her child during pregnancy, sexually (Barzon et al., 2016; D'Ortenzio et al., 2016), or by means of blood transfusions or exposure in a laboratory or healthcare setting (Lazear and Diamond, 2016). As with other arboviruses, about 80% of ZIKV-infected subjects are asymptomatic; symptomatic subjects most frequently experience flu-like syndrome, an itchy maculopapular rash and arthritis or arthralgia, but some cases of retro-orbital pain, headache, myalgia and vomiting have also been observed (Giovanetti et al., 2016). ZIKV can cause severe neurological complications such as Guillain-Barré syndrome and microcephaly in infants born to ZIKV-infected women, as demonstrated by the presence of the virus in the brain, placenta or serum of aborted fetuses and newborns with microcephaly (Sebastian et al., 2017). Since the early 1950s,

101 ZIKV infection outbreaks have been reported in tropical Africa, south-eastern Asia and the Pacific
102 islands. The virus was first isolated in Malaysia in 1969 and, later, in Indonesia (Marchette et al.,
103 1969). In 2007, it caused the first large and well-characterised outbreak on Yap Island, a part of the
104 Federated States of Micronesia (Duffy et al., 2009), and this was followed by a major epidemic in
105 French Polynesia in 2013-2014 that affected more than 28,000 people (11% of the population)
106 (Musso et al., 2018) after which it spread to other neighbouring islands in the south Pacific.

107 In early 2015, the autochthonous transmission of ZIKV in the northeastern part of Brazil was the
108 first reported description of the infection in the Americas (Calvet et al., 2016) and, by the end of the
109 same year, ZIKV activity had expanded into at least 14 Brazilian states (<https://www.paho.org>).
110 Other indigenous cases of ZIKV infection were detected in Colombia, Suriname, Paraguay and
111 Venezuela in south America; Guatemala, El Salvador and Mexico in Central America, and
112 Martinique and Puerto Rico in the Caribbean (Garcia-Luna et al., 2018). In early 2016, local
113 outbreaks were confirmed in Guyana, Ecuador, Bolivia, Peru, Nicaragua, Curacao, Jamaica, Haiti,
114 Santo Domingo and other Caribbean islands. In the first half of the same year, the virus was also
115 detected in Argentina and Cuba and, finally, in the spring 2016, it reached the United States
116 (Florida) (WHO data available at <http://www.who.int/emergencies/zika-virus/history/en/>). The
117 epidemics started to decline in various American countries in the second half of 2016 and, although
118 small local outbreaks were still being reported in 2017, their incidence was greatly reduced.
119 Ultimately, a total of 48 countries in the Americas had more than 540,000 autochthonous cases
120 (more than 200,000 in Brazil) and there were about 2,610 reported congenital infections (2,300 in
121 Brazil) [WHO-PAHO: Regional Zika Epidemiological Update (Americas) August 25, 2017,
122 available at <https://www.paho.org>].

123 ZIKV has now become endemic not only in South America and Caribbean, but also in several
124 Pacific islands (American Samoa, the Federated States of Micronesia, Fiji, Marshall Islands, New
125 Caledonia, Samoa and Tonga) (Calvez et al., 2018). In addition, there have been an increasing

number of travel-related cases in non-endemic countries such as Australia, Belgium, Canada, China, France, Portugal, Spain, Switzerland and The Netherlands (De Smet et al., 2016). Several authors have attempted to study the dynamics of Zika virus infection through discrete phylogeographical analysis (Boskova et al., 2018; Faria et al., 2017; Giovanetti et al., 2016; Liang et al., 2017; Metsky et al., 2017; Pettersson et al., 2018). In addition to classical discrete phylogeographical methods, new continuous diffusion models based on Brownian or random walk diffusion models have been developed that can infer ancestral states on the basis of the coordinates of a two-dimensional space identifying the tips of the tree (sampling location). These models allow a more realistic reconstruction of spatial movements because, unlike discrete models, they do not necessarily need the ancestral location to be represented in the sampling location set (Lemey et al., 2010). The differences between discrete and continuous phylogeographic models have been efficiently described in previous reviews (Bloomquist et al., 2010; Faria et al., 2017). The aim of this study was to infer the origin and dispersion routes of ZIKV in the world using a classical discrete method and to reconstruct the recent epidemic in the Americas using a continuous phylogeographical method to better describe the local spread of ZIKV and to make hypothesis about the eco/epidemiology of the virus.

143

144 **2. MATERIALS AND METHODS**

145 **2.1 Patients and datasets**

146

The study was conducted using 135 complete viral genome sequences retrieved from public databases. These sequences were derived from mosquitoes ($n = 21$), a sentinel rhesus monkey ($n = 6$), and human samples ($n = 108$). The sequences were isolated in various countries of the world and

150 retrieved from GenBank (at <http://www.ncbi.nlm.nih.gov/genbank/>); only the sequences with a
151 known location and sampling date were considered.

152 The sampling period ranged from 1947 to 2016, and the sampling locations ranged from the Central
153 African Republic (CF, n=4) to Nigeria (NG, n=2), Senegal (SN, n=10), Uganda (UG, n=6), south-
154 east Asia (SEA, including Malaysia n=4, Cambodia n=2, Thailand n=3, and Singapore n=2), French
155 Polynesia (FP, n=11), Mexico (MX, n=8), Honduras (HN, n=6), Guatemala (GT, n= 3), Panama
156 (PA, n=4), Venezuela (VE, n=8), Colombia (CO, n=4), Ecuador (EC, n= 2), Brazil (BR, n=28),
157 Suriname (SR, n=2), the Western Pacific (WP, including American Samoa n=6, Tonga n=1, the
158 Philippines n=1, Micronesia n=1), the Antilles (ANT, including Puerto Rico n=4, Haiti n=2,
159 Dominican Republic n=8, Martinique n=1, Cuba n=2). Two of the 28 Brazilian samples were from
160 tourists returning to Italy who became infected in Bahia State.

161 The sequences were selected on the basis of the following criteria: i) they had to have been
162 published in peer-reviewed journals; ii) their non-recombinant subtype assignment had to be
163 certain; and iii) the city/state of origin and year of sampling had to be known and clearly established
164 in the original publication. The origins and characteristics of the Zika strains dataset are
165 summarised in **Supplementary Table 1**.

166

167 **2.2 Ethics Statement**

168 Informed consent was obtained according to Italian law (art.13 D.Lgs 196/2003) as well as approval
169 of the Institutional review board of Fondazione IRCCS Policlinico San Matteo on the use of
170 residual biological specimens (IRB Protocol 20100000348).

171

172

173 **2.3 Whole genome characterization by means of next-generation** 174 **sequencing** 175

176 The whole ZIKV genome sequence of one human isolate (an Italian semen sample from a traveller
177 coming from Bahia State in Brazil) was previously obtained by Sanger methodology and deposited
178 in GenBank database with the accession number KY003154. Subsequently, in our laboratory, we
179 amplified the whole genome by using the sequence-independent single-primer amplification (SISPA
180 method) and re sequenced it by NGS method (Djikeng et al., 2008).

181 ZIKV was isolated on Vero E6 cells; the RNA was prepared by extracting it from the cell culture
182 supernatant and then it was reverse-transcribed using the random primer FR26RV-N
183 (5'GCCGGAGCTCTGCAGATATCNNNNNN3') at a concentration of 10 µM. Viral cDNA was
184 denatured at 94°C for three minutes, and chilled on ice for two minutes. Five units of Klenow
185 fragment (New England Biolabs, Ipswich, MA) were directly added to the reaction to perform the
186 second strand cDNA synthesis. The incubation was carried out at 37°C for one hour, and at 75°C
187 for 10 minutes.

188 Next, 5 uL of double-stranded DNA were added to a PCR master mix containing 5 µL of 10x
189 AccuPrime PCR buffer I, 0.2 µL of AccuPrime Taq DNA Polymerase high fidelity, 4 uL of 10 µM
190 FR20RV (5'GCCGGAGCTCTGCAGATATC3') and 35.8 µL of water. The incubation was
191 performed under the following thermal conditions: 94°C for two minutes, 40 cycles of 94°C for 30
192 seconds, 55°C for one minute and 68°C for three minutes.

193 The PCR product was purified and quantified using a TECAN plate reader. The sample was diluted
194 to an initial concentration of 0.2 ng/µL in accordance with the Illumina protocol, and 1 ng was used
195 for the library preparation (Nextera XT sample preparation Kit, Illumina Inc., San Diego,
196 California, USA).

197 Genomic libraries were sequenced on the Illumina MiSeq platform (Illumina, Inc.) with 2x151 base
198 pairs paired-end runs. Finally, we evaluated the obtained reads for sequence quality and read-pair
199 length using FastQC ver. 0.11.5

200 The reads were assembled using Geneious software v. 11.1.5 (Biomatters , New Zealand) and re-
201 sequencing analysis was performed with the reference virus (KY003154).

202

203 **2.4 Recombination detection**

204

205 In order to identify recombinant strains and exclude them from the analysis, we used the RDP4
206 package, which allows the identification of potential recombinant sequences and their parents
207 (major and minor). It uses seven different methods: RDP (Martin et al., 2015), BOOTSCAN
208 (Martin et al., 2005), CHIMAERA (Posada and Crandall, 2001), SISCAN (Gibbs et al., 2000),
209 GENCONV (Padidam et al., 1999), 3SEQ (Boni et al., 2007) and MAXCHI (Smith, 1992), each of
210 which has a highest acceptable p value of 0.05 and Bonferroni's correction for multiple comparisons
211 The sequences indicated as being recombinant by at least three of these methods were excluded
212 from further analysis.

213 We also screened our alignment using Genetic Algorithm Recombination Detection (GARD)
214 software in order to detect any sequences involved in putative recombinations, and define the
215 number and location of breakpoints (Kosakovsky Pond et al., 2006).

216

217 **2.5 Likelihood mapping**

218

219 The phylogenetic signal of the complete genome dataset was investigated by means of the
220 likelihood mapping (LM) analysis of 10,000 random quartets generated using TreePuzzle (Strimmer
221 and von Haeseler, 1997). Groups of four randomly chosen sequences (quartets) were evaluated and,
222 for each quartet, the three possible unrooted trees were reconstructed using the maximum likelihood

223 approach under the selected substitution model that was the General Time Reversible(GTR) with
224 gamma distributed rates among sites. The posterior probabilities of each tree were then plotted on a
225 triangular surface so that fully resolved trees fell into the corners, and the unresolved quartets in the
226 centre of the triangle (a star-tree). When using this strategy, if more than 30% of the dots fall into
227 the centre of the triangle, the data are considered unreliable for the purposes of phylogenetic
228 inference.

229 **2.6 Phylogenetic reconstruction**

230

231 The sequences were aligned using ClustalX software (Jeanmougin et al., 1998) followed by manual
232 editing using Bioedit software v. 7.2.6, and the best fitting nucleotide substitution model was tested
233 by means of the hierarchical likelihood ratio test (LRT) implemented in J Modeltest software
234 (Posada, 2008). The selected model was GTR with gamma distributed rates among sites.

235 The phylogeny of the complete genome was reconstructed using a maximum likelihood approach
236 and the new hill-climbing algorithm implemented in PhyML v.3.0. The reliability of the observed
237 clades was established on the basis of internal node bootstrap values of $\geq 70\%$ (after 200 replicates)
238 (Guindon et al., 2010).

239 The phylogeny of the complete genome was also reconstructed using a Bayesian Markov Chain
240 Monte Carlo (MCMC) method (Beast v. 1.8.4 freely available at <http://beast.bio.ed.ac.uk>). The
241 reliability of the observed clades was established on the basis of posterior probabilities values with
242 significance levels of ≥ 0.7 .

243 The evolutionary rates were estimated under strict and relaxed (with an uncorrelated log normal rate
244 distribution) clock conditions.

245 As coalescent priors, three parametric demographic models of population growth (constant size,
246 exponential growth and logistic growth) the Bayesian SkyGrid, the Bayesian skyline plot (BSP) and
247 the GMRF Bayesian Skyride were compared (Drummond et al., 2005). The best fitting models were
248 selected using the BF implemented in Beast. In accordance with Kass and Raftery, the strength of

the evidence against H_0 was evaluated as $2\ln BF < 2$ = no evidence; $2-6$ = weak evidence; $6-10$ = strong evidence, and >10 = very strong evidence. A negative $2\ln BF$ indicates evidence in favour of H_0 . Only values of ≥ 6 were considered significant (Kass and Raftery, 1995). We also used path sampling (PS) and stepping stone sampling (SS) to improve the accuracy of model selection (Baele et al., 2012). The chains were run for 250 million generations until reaching convergence and sampled every 25000 steps. Convergence was assessed by estimating the effective sampling size ($ESS = >200$) after a 10% burn-in, using Tracer software version 1.6 (<http://tree.bio.ed.ac.uk/software/tracer/>). All of the parameters had an ESS of >200 . Uncertainty in the estimates was indicated by 95% highest posterior density (95% HPD) intervals. The TMRCA estimates were expressed as the median and 95% HPD years before the most recent sampling date, which corresponded to 2016 in this study.

261

262 **2.7 Root-to-tip regression analysis**

263

In order to verify the correlations between time and genetic distances and identify the correct root under the hypothesis of proportionality between them, we used Tempest software and the ML tree to make a root-to-tip regression analysis (Rambaut et al., 2016).

267

268 **2.8 Bayesian phylogeographic analyses**

269

270 **2.8.1 Discrete phylogeographic analysis**

271

An improvement to Beast allows an ancestral reconstruction of discrete states in the Bayesian framework described above in which the spatial diffusion of the time-scaled genealogy is modelled as a continuous-time Markov chain process over discrete sampling locations. A Bayesian stochastic

275 search variable selection (BSSVS) approach, which allows the exchange rates in the CTMC to be
276 zero with some prior probability, was used in order to find a minimal (parsimonious) set of rates
277 explaining the diffusions in the phylogeny. Comparing the posterior to prior odds that individual
278 rates are zero provides a Bayes factor test to identify the rates contributing to the migration
279 pathway, which were calculated as described elsewhere. Rates yielding a BF of >3 were considered
280 significant (Lemey et al., 2009).

281 The obtained trees were summarised in a maximum clade credibility tree using the Tree Annotator
282 program included in the Beast package, choosing the tree with the maximum product of posterior
283 probabilities (maximum clade credibility: MCC) after a 10% burn-in. The most probable location of
284 each node was highlighted by labelling the branches with different state colours. In order to
285 visualize diffusion rates over time, it is possible to convert the location-annotated MCC tree to a
286 keyhole mark-up language file (KML) suitable for viewing with georeferencing software.

287 In order to visualize diffusion rates over time, it is also possible to render the location-annotated
288 MCC tree to a GeoJSON data format suitable for viewing with georeferencing software. The new
289 SPREAD3 analysis tool was used, the MCC tree was converted to a JavaScript object notation
290 (JSON) file and the visualization was rendered using a Data Driven Document (D3) library
291 (Bielejec et al., 2016).

292 **2.8.2 Continuous phylogeographic analysis**

293

294 In order to study the spread of ZIKV in more detail, a continuous space phylogeographical analysis
295 was made using American isolates.

296 American ZIKV epidemics were investigated in continuous space using Beast v. 1.8.4. The
297 unknown coordinates were estimated under a strict Brownian diffusion model, and compared with
298 two relaxed random walk (RRW) models relaxing the diffusion rate constancy assumption that
299 respectively assumed the gamma and Cauchy distribution of diffusion rates over the phylogeny
300 (Lemey et al., 2010). Bayes factor comparisons of the models were made by estimating marginal

301 likelihood using path sampling (PS) and stepping stone approaches (Baele et al., 2012). The
302 phylogeny was spatially projected and converted into KML in order to visualize dispersal over time.
303 Uncertainties in the ancestral location estimates were represented by KML polygons delimiting the
304 high-probability regions.

305 **3. RESULTS**

306

307 **3.1 Illumina paired-end sequencing**

308

309 The raw data reads with quality value $QV > 20$ were filtered by excluding contaminants such as
310 adapters, the ambiguous “N” nucleotides and low-quality sequences using trimming options
311 implemented in Geneious software. After trimming the raw data, a total of 178846 filtered clean
312 reads were obtained with a 76023X coverage.

313 **3.2 Recombination analysis**

314

315 Both the GARD and RDP programs confirmed the presence of recombination in the final
316 alignment. In particular, RDP analysis (Supplementary Table 2) detected four genomes showing a
317 total of 12 significant recombination events corresponding to isolates from Senegal (12SN@68,
318 13SN@97, 14SN@01 and 15SN@01). For this reason, these sequences were removed from the
319 definitive data set.

320 **3.3 Likelihood mapping analysis**

321

322 The presence of phylogenetic noise was investigated using LM analysis. The complete genome data
323 set gave satisfactory results as 7.8% of the dots fell into the central area of the triangles and 87.4%
324 at the corners, thus suggesting that the alignment contained sufficient phylogenetic information
325 (Supplementary Fig. 1).

3.4 Phylogenetic analysis

ML analysis of the whole genomes showed two statistically supported clades (bootstrap=1000) corresponding to the previously described African (AF) and Asian (AS) clades (**Fig. 1**). Two different sub-clades could be distinguished within the African clade: the first (eastern central African, EC) sub-clade included the original 1947 Ugandan isolates, four genomes from the Republic of Central Africa obtained between 1968 and 1980, and two Senegalese isolates obtained in 2001 (each of them grouping together on a geographical basis); the second (Western African, W) sub-clade encompassed the majority of the Senegalese isolates obtained between 1968 and 1997 and two Nigerian strains obtained in 1968, all significantly segregating on the basis of their geographical origin.

The Asian/Pacific Ocean clade (AS) included two geographically distinct monophyletic groups: the first including all of the isolates from Malaysia 1966 (MY), and the second and largest group (AWP) including all of the other Asian, Western Pacific Ocean, Polynesian and American isolates connected by a long branch (indicating a bottleneck) to the Malaysian clade. The American sub-clade was statistically sustained (bootstrap=1000). Analysis of the single genes highly supported these clades and sub-clades (**Fig. 1**).

3.5 Root-to-tip regression analysis

Analysis of the unrooted ML tree of the entire dataset without the recombinant sequences, showed a very strong association between genetic distances and sampling dates ($R^2=0.93$, correlation coefficient=0.97), thus confirming the suitability of the dataset for molecular clock analysis. Tempest also located the best tree root within the branch connecting the African and Asian clades. Interestingly, separate analysis of each gene (**Table 1**) showed similar results in all the datasets, with the weakest temporal signal being obtained using the Membrane dataset ($R^2=0.5$).

352

353 **Table 1.** Comparison between regression analysis and phylogeographic analysis for each single
 354 *Zika virus* gene.

355

	ROOT-TO-TIP					BAYESIAN ANALYSIS	
	SLOPE (*10 ⁻³)	tMRCA ¹	CORRELATION COEFFICIENT	R SQUARED	RESIDUAL MEAN SQUARED	E.R ² MEAN (95% HPD ³ LOWER-UPPER)	tMRCA ¹
CAPSID	1.19	1878.3	0.97	0.94	2.97*E ⁻⁵	1.03 (0.42-1.6)	1936.5
MEMBRANE	1.02	1885.4	0.73	0.54	3.2*E ⁻⁴	2.2 (1.06-3.5)	1944.8
ENVELOPE	0.85	1802.2	0.91	0.84	4.98*E ⁻⁵	1.7 (0.95-2.4)	1940.6
NS1	0.45	1806.1	0.92	0.84	1.32*E ⁻⁵	1.4 (0.92-2)	1944.3
NS2	1.18	1811.3	0.95	0.9	5.48*E ⁻⁵	1 (0.46-1.6)	1924.9
NS3	0.67	1863.2	0.97	0.94	9.99*E ⁻⁶	0.85 (0.53-1.2)	1924.7
NS4	0.59	1826.9	0.95	0.9	1.33*E ⁻⁵	0.86 (0.58-1.3)	1910.1
NS5	0.63	1835.4	0.97	0.93	1.03* E ⁻⁵	0.93 (0.63-1.3)	1913.6

¹ tMRCA: time of the most Recent Common Ancestor

² E.R: Evolutionary Rate

³ HPD: Highest posterior density, substitutions/site/year (*10⁻³)

356 **3.6 Evolutionary rates, tMRCA estimates and Bayesian** 357 **phylogeography**

358

359 The evolutionary rates, tMRCAs and phylogeography were co-estimated using a Bayesian
 360 framework implemented in Beast (v. 1.8.4). The comparison by Bayes factor of the marginal
 361 likelihoods obtained by applying a strict or relaxed molecular clock under five different coalescent
 362 models (2lnBF GMRF Bayesian Skyride vs BSP = - 811.92; 2lnBF constant vs BSP = - 68.31;
 363 2lnBF exponential growth vs BSP = -1092.56; 2lnBF Bayesian Skygrid vs BSP = - 748.1); under a
 364 log-normal relaxed clock (2lnBF strict vs relaxed clock = -466.62) showed that the favoured
 365 models were the relaxed molecular clock with uncorrelated log-normal rate distribution and the
 366 Bayesian skyline plot, the less stringent demographic model. The same model has been confirmed
 367 by using path sampling model selection (PS) and stepping stone sampling (SS) model selection
 368 (**Table 2**). Under these conditions, we estimated a mean substitution rate for the entire viral genome

369 of 8×10^{-4} (95% HPD $6.4\text{-}9.7 \times 10^{-4}$) subs/site/year (**Table 1** shows the mean estimates for each
370 single gene).

371 The tree-root tMRCA (**Fig. 2**) was estimated to be an average of 114.2 (95% HPD 84-148) years
372 before the present, corresponding to the year 1902 (95% HPD 1868-1932). The MRCA of the
373 eastern Central African clade was placed in 1930 (95% HPD=1918-1940), whereas that of the West
374 African sub-clade originated later, in 1954 (95% HPD=1944-1963). The first Asian node
375 (corresponding to the Malaysian group) dated back to 1958 (95% HPD 1951-1965), and was
376 connected by a long branch to the largest AWP sub-clade whose tMRCA dated back to 1998 (95%
377 HPD 1993-2003). A further highly significant sub-clade (pp=1) included all of the American
378 strains, which had an estimated mean tMRCA of 3.1 years ago (95% HPD 2.8-3.5), corresponding
379 to the year 2013 (2012-2014). In general, the isolates of this AWP clade also tended to segregate
380 significantly on the basis of their geographical origin (**Table 3**).

381 Analysis of migration flows showed 19 (95% HPD=17-21) non-zero rates between different
382 localities, all of which were significant at BF analysis (BF>3). The suggested dispersion pathway
383 summarized in **Fig. 3** showed that the currently circulating ZIKV strains shared a common ancestor
384 that existed in eastern Central Africa in the first decades of the 1900s, and spread to Western Africa
385 in the 1950s. In the same period, it spread to Asia (Malaysia) and the Asian strain reached the
386 Pacific islands at least twice: in the first decade of the 2000s and in 2012, when it spread to French
387 Polynesia. Finally, from French Polynesia, it reached Brazil in 2013 to start a flow that
388 subsequently reached a several number of central and south American regions.

389 Considering the discrete phylogeographic tree and limiting the analysis to the Asian clade without
390 the Malaysian strains, we observed a total of 14 highly significant ($0.9 < \text{pp} < 1$) monophyletic groups
391 including more than two isolates (a median of five isolates for clade, range 3-9). The clades were
392 strongly defined on a geographical basis (**Supplementary Table 3**): five were Brazilian clades
393 (three including only Brazilian isolates, and two including mixed isolates from Brazil and Ecuador
394 or Italy); two included Venezuelan isolates in one clade with isolates from Colombia and the other

395 with one Dominican sequence; two were pure Central American clades (one from Mexico and the
 396 other from Panama) and two others were from the Caribbean (one pure from Puerto Rico and one
 397 mixed from Cuba and Santo Domingo, with one strain from Mexico). The earliest tMRCA was that
 398 of the French Polynesia clade, followed by the Brazilian clade, the Central American clades and,
 399 finally the clades from Antilles and Venezuela.

400 In order to reconstruct the spread of ZIKV in the new world in more detail and avoid the limitations
 401 caused by their arbitrary grouping into discrete localities, we used a continuous phylogeographic
 402 model based on the geographical coordinates of the sampling localities. This analysis only
 403 considered data subset of American isolates. Comparison of the strict Brownian diffusion model
 404 (assuming a homogeneous diffusion rate over the phylogeny) with two RRW models (assuming
 405 different diffusion rates on each branch of the tree) by the BF test showed that a log-normal RRW
 406 diffusion rate fitted the data better than the other models (Cauchy distribution RRW *vs* homogenous
 407 BD: $2\ln BF = 639.34$ by PS and 175.78 by SS; Cauchy distribution RRW *vs* log-normal RRW:
 408 $2\ln BF = 1335.32$ by PS and 10.64 by SS). On the basis of this continuous phylogeographical
 409 reconstruction, the tree root was placed between the coordinates -41.13 E of longitude and -9.53 N
 410 of latitude, corresponding to a location in the state of Bahia in north-east Brazil, close its border
 411 with the two other states of Pernambuco and Piauí (see the animated visualization in
 412 **Supplementary Video 1**). Two high probability (80% HPD) regions (**Fig. 4**) were identified almost
 413 simultaneously at the beginning of the epidemic (2013): the first (A in Fig. 4, panel 1) was a large
 414 ellipse with a major axis of about 1700 km and a minor axis of about 700 km that included the tree-
 415 root and encompassed north-east Brazil (the nine states of Alagoas, Bahia, Ceará, Maranhão,
 416 Paraíba, Pernambuco, Piauí, Rio Grande do Norte, and Sergipe); the second (B in Fig. 4, panel 1)
 417 was a smaller circle with a diameter of about 400 km, surrounding the metropolitan area of Rio de
 418 Janeiro in south-east Brazil. These two areas were the origin of different migration flows,
 419 corresponding to the branches and clades highlighted in the continuous phylogeographic tree (Fig.
 420 4, panels 1-4, and **Fig. 5**). Area A gave rise to two main migration pathways: the first

(corresponding to clade A1 in the tree) initially spread across north-east Brazil and subsequently reached Santo Domingo (2015), Cuba and Ecuador (2016); the second (corresponding to clade A2, and not significantly segregated from the root of the tree) first reached the island of Haiti (2014) and then spread to Venezuela (2016). Area B was the origin of three main migratory flows: the first (B1) reached Suriname in 2013, and then proceeded towards Porto Rico (2015); the second (corresponding to clade B2) spread to an uncertainty region encompassing Panama and Colombia in 2013 and subsequently (2015/2016) dispersed east towards the Caribbean (Martinica, Santo Domingo) and Venezuela, and north towards Mexico; and the third (B3) reached Honduras and Guatemala before spreading throughout Central America and reaching Mexico in the last year (2016).

Table 4 shows the geographical coordinates of the localities and tMRCA estimates for each of the main clades (root, A1-2, B1-3). The estimated average tree-root tMRCA was 3.7 years ago, corresponding to October 2012 (95% HPD December 2011 to August 2013), and the estimated average tMRCAs of the main clades were between March (clade A2) and August 2013.

Within a few months, ZIKV had spread to the entire continent. It followed a mainly north-westerly pathway at the beginning of the epidemic but, between late 2014 and the beginning of 2015, went in various directions (along an east/west axis between central America and the Caribbean, and even in a south-easterly direction), thus indicating wider viral dispersion throughout the region. The estimated overall diffusion rate was as much as 760 km/year (between about 600 and 900 km/year).

447 **Table 2. Model selection using Path Sampling and Stepping Stone Sampling.**

CLOCK	MODELS	COMPLETE GENOME PS ¹	COMPLETE GENOME SS ²
Strict	Constant	-37209,53	-37259,37
Strict	Exponential	-37204,45	-37271,32
Strict	Skygrid	-37208,28	-37265,18
Strict	Skyline	-37173	-37235,26
Strict	Skyride	-37417,47	-37499,97
UCLN ³	Constant	-37148,91	-37218,2
UCLN ³	Exponential	-37153,79	-37225,78
UCLN ³	Skygrid	-37151,84	-37196
UCLN ³	Skyline	-37100,32	-37163,92
UCLN ³	Skyride	-37362,07	-37411,22

¹ Path Sampling
² Stepping stone sampling
³ Uncorrelated log normal

448
449

450 **Table 3.** Estimated times of the most recent common ancestors (tMRCAs) of the main clades and
451 credibility intervals (95% HPD), with calendar years, most probable locations, and state posterior
452 probabilities (spp) of the 131 complete genomes of *Zika virus*.

453

		tMRCA ¹			YEARS			location	stpp ⁴
Nodes	pp ²	MEAN	95% HPD ³ lower	95% HPD ³ upper	MEAN	95% HPD ³ Lower	95% HPD ³ upper		
Root	1	114,2	148,44	84,08	1902	1868	1932	UG	0,26
African	0,71	98,6	114,18	83,61	1917,4	1901,8	1932,4	UG	0,34
Eastern Central	0,97	86,1	97,77	76	1929,9	1918,2	1940	UG	0,51
Western	1	61,6	71,57	53,26	1954,3	1944,43	1962,7	NG	0,57
Asian	1	57,7	65,15	51,43	1958,2	1950,85	1964,6	SEA	0,81
American	1	3,1	3,46	2,77	2012,9	2012,5	2013,22	BR	0,83

¹ tMRCA: time of the most Recent Common Ancestor
² pp: posterior probability
³ HPD: highest posterior density
⁴ stpp: state posterior probability

454
455
456
457

Table 4. The main clades, calendar months, most probable locations with mean estimated coordinates, and posterior probabilities (pp) of the 76 complete genomes of *Zika virus*.

CLADES	pp ¹	LONGITUDE	LATITUDE	LOCATION	MONTH	MONTH L ⁵	MONTH U ⁶
Root tree	1	41.126	9.532	BRA_NE ²	October 2012	August 2013	December 2011
A1	0.96	38.8	7.204	BRA_NE ²	June 2013	May 2014	August 2012
A2	0.12	42.056	8.41	BRA_NE ²	March 2013	November 2013	May 2012
B2	0.82	42.457	20.003	BRA_SE ³	April 2013	January 2014	August 2012
B1	0.99	45.678	13.439	BRA_C ⁴	August 2013	May 2014	December 2012
B3	0.93	43.377	21.865	BRA_SE ³	May 2013	February 2014	October 2012

¹ Posterior probability

² North-east Brazil

³ South-east Brazil

⁴ Centre Brazil

⁵ Lower

⁶ Upper

4. DISCUSSION

A number of authors have recently attempted to estimate the evolutionary dynamics of ZIKV in the Americas with the aim of reconstructing the most probable origin of the epidemic, the time of its entry into the Americas, and the diffusion pathways that led to its spread across the continent in such a short time. Some of these studies used partial coding sequences (Giovanetti et al., 2016; Liang et al., 2017), and others whole viral genomes (Faria et al., 2017; Metsky et al., 2017). Some authors included in their analyses all of the isolates available in public databases obtained in over 70 years since the first isolates in 1947 (Giovanetti et al., 2016; Liang et al., 2017), whereas others concentrated the study only on American (Boskova et al., 2018; Faria et al., 2017) or Asian isolates (Pettersson et al., 2018).

474 In this study we reconstructed the spatiotemporal dynamics of ZIKV at a global and a local scale by
475 using, for the first time, two different phylogeographic approaches: a discrete and a continuous
476 diffusion model.

477 In order to investigate the phylogenetic information contained in partial genes, we made a root-to-
478 tip regression analysis that showed a sufficient temporal structure ($R^2 = 0.93$) in the whole genome
479 dataset, whereas the results of analyses of the individual gene datasets were ambiguous in terms of
480 locating the best-fitting root position and estimates of the evolutionary rates and root - tMRCA. For
481 these reasons, the dataset was also analysed for the presence of recombination, which revealed four
482 recombinant genomes isolated in Senegal in different years (1968, 1997 and 2001), carrying a total
483 of 12 recombination breakpoints. Homologous recombinations in ZIKV has been previously
484 described (Faye et al., 2014; Han et al., 2016) also in other flaviviruses (Simon-Loriere and Holmes,
485 2011), and may explain some of the discrepancies in dating the origin of the virus, particularly
486 when this is based on partial genomes.

487 However, although it therefore seems to be essential to use whole genomes and exclude
488 recombinant sequences in order to obtain unbiased inferences, complete genomes are relatively
489 scarce because of the difficulties in performing extensive genome sequencing (Metsky et al., 2017).
490 The low level of viremia frequently requires preparatory cultural enrichment of the virus, and so we
491 have developed a protocol for the whole genome sequencing of ZIKV that was implemented on the
492 Illumina NGS platform. The full genome of an Italian sample of human semen that has been
493 previously characterised using a Sanger-based method was further analysed using our NGS
494 protocol, revealing over 99% identity with KY003154. Our sequence was aligned with 130 other
495 complete ZIKV genomes for which the sampling dates and locations were known that were
496 retrieved from public databases at the time the analysis began.

497 Our findings indicate that the mean evolutionary rate of the viral genome was between 8.5 and
498 22×10^{-4} , similar to the estimates previously obtained by other authors (evolutionary rates ranging
499 from 6 to 13×10^{-4}) (Liang et al., 2017). On the basis of this evolutionary rate, we estimated a tree-

500 root tMRCA (varying from 84 to 148 years), corresponding to the late 19th and early 20th centuries.

501 In agreement with Liang (Liang et al., 2017), who only studied partial genes, our data suggest that

502 the Zika virus emerged more than 100 years ago as an endemic/epidemic infection, probably after a

503 period of sylvatic circulation in the rainforests of the east Africa. The virus then spread to West

504 Africa in the 1950s and, a few years later (1958s), to south-east Asia (Malaysia). After a long period

505 of about 40 years without any new isolation, causing a bottleneck effect (highlighted by the long

506 branch of the tree connecting the older Malaysian strains to the new SEA isolates), the virus

507 reappeared in the Pacific islands (Micronesia) in the early 2000s, and caused the first significant

508 outbreak on the Island of Yap in 2007. The virus again left south-east Asia and spread to French

509 Polynesia where it caused the largest outbreak ever recorded before that time. Our estimated

510 tMRCA showed that the virus had been present in French Polynesia at least since the second half of

511 2012 even if the first recordable cases were in late 2013 (between the 41st week of 2013 and the 14th

512 week of 2014). Our data confirm that ZIKV reached French Polynesia from south-east Asia not

513 from the Island of Yap, as suggested by others (Pettersson et al., 2018; Weaver et al., 2016).

514 Finally, it moved from French Polynesia to the Americas (Brazil) and spread throughout the

515 continent within a few years. Interestingly, the skyline analysis showed an exponential increase in

516 the effective number of infections from early 2013 to late 2014, which corresponds with the spread

517 of the virus in the Western hemisphere, according to our reconstruction (**Supplementary Fig. 2**).

518 The added value of this study comes from the use of a continuous diffusion model to reconstruct the

519 spread of ZIKV in the Americas. Discrete phylogeography requires the grouping of isolates into

520 categories that can be based on political-administrative boundaries, or other geographically or

521 epidemiologically homogenous areas/populations. However, the grouping is frequently arbitrary

522 and often lacks precision in reconstructing flow rates between different spatial areas, in particular

523 when there are ambiguities in assigning the isolates to one group or another. Continuous

524 phylogeography allows these limitations to be overcome by identifying isolates on the basis of

525 geographical coordinates corresponding to the latitude and longitude of the sampling location. We

526 deduced these coordinates on the basis of the patients' data available, and the use of diffusion
527 models made the reconstruction of the spread of the epidemic in the Americas more precise by
528 allowing ancestral sequences to reside at any location in a continuous bi-dimensional space
529 whereas, in the case of discrete approaches, there is no way to infer the ancestral location from the
530 tree if it is not present in the sampled location set.

531 Our continuous phylogeographical reconstruction allowed an estimate of the possible geographical
532 coordinates of the entry location and an estimated 80% probability region covering the entire north-
533 east of Brazil, thus suggesting that ZIKV entered Brazil in late 2012. Just two years later, in early
534 2015, cases of a “dengue-like syndrome”, that probably represented the first cases of the ZIKV
535 epidemic, were reported in two cities: Natal (in the State of Rio Grande del Norte) and Camaçari (in
536 the metropolitan area of Bahia) (Campos et al., 2015; Zanluca et al., 2015). Both cities are included
537 in the 80% probability region of our estimated tree-root even if the analysed samples did not include
538 any coming from these places which underlines the importance of using these models.

539 Interestingly, the analysis also identified a second high probability region surrounding Rio de
540 Janeiro at the very beginning of the epidemic (2013) as a further initial area of viral dissemination
541 across the Americas. A recent study using a data-driven stochastic and spatial epidemic model
542 considering the period between April 2013 and June 2014 estimated the arrival of ZIKV in the
543 Americas in August 2013 in Rio de Janeiro or north-east Brazil, where mosquito density and DENV
544 transmission is highest (Zhang et al., 2017). This is in line with our spatial reconstruction, although
545 our tree root tMRCA is several months earlier (October 2012, with an upper estimate of August
546 2013). Furthermore, a recent molecular survey identified the presence of ZIKV RNA in samples
547 collected in March-May 2013 in Tijuca (in the metropolitan area of Rio de Janeiro) from patients
548 with acute febrile syndrome negative for DENV RNA (Passos et al., 2017), and other phylogenetic
549 studies based on human and entomological samples support the presence of ZIKV in Brazil in late
550 2012 and early 2013 (Ayllon et al., 2017; Metsky et al., 2017).

Two waves of the ZIKV epidemic in Brazil have been hypothesised: an early wave in north-east Brazil and a second in south-east Brazil (Zhang et al., 2017). Our continuous phylogeographical analysis of the American clades identified two clades supporting the hypothesis of a first epidemic wave starting in north-east Brazil in early 2013 that initially spread locally (A1) and reached Haiti (A2). Previous authors have suggested an initial wave of ZIKV in Haiti (Lednicky et al., 2016). The same viral strains were only later exported to Santo Domingo, Cuba and Ecuador.

The second epidemic wave starting in the region surrounding Rio de Janeiro (B) spread over a larger area and give rise to new dispersal centres in Suriname, Panama/Colombia and Honduras/Guatemala from which the infection further spread to the Caribbean islands and Central America. The more extensive spread of the strains originating in region B may have been due to the larger passenger flows related to the importance of Rio as a Brazilian transportation hub (Zhang et al., 2017).

Globally, the virus moved north-west from Brazil to the central-north America at an estimated mean diffusion rate of 760.8 km/year (95% HPD 596-913 km/year) between early 2013 and 2016.

It has recently been suggested that ZIKV may have been imported into the Americas as a consequence of the Confederations Cup held in Brazil in June 2013. It is probable that, during this event, many athletes and supporters from affected areas (French Polynesia) probably travelled through a large area of eastern Brazil stretching from Fortaleza to Rio de Janeiro. An alternative hypothesis is the simultaneous arrival of the same south-east Asian virus in French Polynesia and Brazil (Zhang et al., 2017), but this does not seem to be confirmed by our analysis, which indicates that ZIKV entered Polynesia before reaching Brazil.

One question that deserves to be clarified is the cryptic circulation of the virus in Brazil before the first cases of ZIKV infection were reported in 2013-2015; this also seems to have occurred in Polynesia, where the estimated first entry of the virus is at least one year before the first human cases were reported. The frequency of asymptomatic infections and the presence in the same area of viruses causing infections with similar outcomes (such as *Chikungunya* and *Dengue* viruses) may

577 have masked the initial spread of the virus. Moreover, sylvatic viral circulation can be hypothesised,
578 even if there is only limited evidence of the exposure of non-human primates to Zika virus in the
579 new world (Chiu et al., 2017; Moreira-Soto et al., 2018). Other factors such as the density of the
580 vector population and the seasonality nature of vector abundance (Lana et al., 2014) may also affect
581 the duration of silent circulation.

582 In the absence of any known animal reservoirs and a possible enzootic circulation such as in the
583 case of *West Nile* Virus, which is known to have undergone a latent period of circulation before the
584 appearance of the first human cases (Zehender et al., 2017), the possible role of transmission routes
585 other than *Aedes* mosquito bites should be investigated. It has been reported that the infection can
586 be sexually transmitted due to its persistent presence in the semen of affected males, even if they are
587 asymptomatic. Recent studies have shown that the sexual transmission may contribute to increase
588 the final size and the persistence of the epidemic (Gao et al., 2016). In particular, a recent study
589 showed that up to 47% of ZIKV cases may be due to sexual contacts when *Aedes* mosquitoes are
590 also present (Towers et al., 2016). Sexual transmission and possibly migration of recently infected
591 subjects (Baca-Carrasco and Velasco-Hernandez, 2016; Olawoyin and Kribs, 2018) may have
592 initially contributed to the slow and hidden circulation of the virus before the accumulation of a
593 critical number of infected humans and mosquitoes.

594 This is the first study, to our knowledge, that provides an estimate of the geographic origin and
595 diffusion pathways of ZIKV in America across a continuous space. Moreover, it provides a new
596 ZIKV genome obtained through NGS platform.

597 The main limitation of this study is that it analysed a relatively small number of complete ZIKV
598 genomes, although it must be remembered that the databases included only 135 complete genomes
599 at the time the study was started. It is not a limitation in terms of coalescent theory, which considers
600 small samples of whole populations (Griffiths and Tavaré, 1994) but, as it may be a problem for
601 phylogeographical studies in terms of sampled locations, we partially compensated this by using
602 continuous phylogeography. Unfortunately, the scarcity of publicly available sequences is

603 essentially due to the difficulty in detecting the virus in samples and the need for cultures in order to
604 have sufficient material (Metsky et al., 2017)

605 Our data underline the importance of genome characterisation and the use of phylogeography in the
606 surveillance of emerging infections.

607 **Acknowledgements**

608
609 This work was partially financed by the “NANOMAX” Bandiera project

610 (2013–2015) funded by Italian Ministry for Education, University and Research (grant number
611 G42I12000180005) to GZ.

612 This study was also partially supported by funds from Lombardy Region and grant from
613 the Ministero della Salute, Ricerca Corrente Fondazione Istituto Di Ricovero e Cura a
614 Carattere Scientifico Policlinico San Matteo, grant no. 80206 to EP.

615

Figures legend

Fig. 1. Maximum likelihood tree of the 133 *Zika virus* complete genome sequences. The significant posterior probabilities ($pp \geq 0.7$) of the corresponding nodes have been coloured in red and the main clades have been highlighted. The scale bar indicates 2% of nucleotide divergence.

Fig. 2. Phylogeographic analysis of 131 *Zika virus* isolates in the world. The branches of the maximum clade credibility (MCC) tree are coloured on the basis of the most probable location of the descendent nodes (ANT= Antille; BR= Brazil; CF= Central African Republic; CO= Colombia; EC=Ecuador; GT= Guatemala; HN= Honduras; MX= Mexico; NG= Nigeria; PA= Panama; PF= French Polynesia; SEA= South Eastern Asia; SN= Senegal; SR= Suriname; UG= Uganda; VE= Venezuela; WP= Western Pacific). The numbers on the internal nodes indicate posterior probabilities \geq to 0.7, and the scale at the bottom of the tree represents calendar years. The main geographical clades are highlighted.

Fig. 3. Spatio-temporal dynamics of the *Zika virus* epidemic in the world. The figure summarises the most significant migration links in the involved areas.

Fig. 4. Spatio-temporal dynamics of the *Zika virus* epidemic in the Americas. The figure summarises the most significant migration links in the involved area.

Fig. 5. Phylogeographical analysis of 76 *Zika virus* isolates in the Americas. The numbers on the internal nodes indicate posterior probabilities of >0.7 , and the scale at the bottom of the tree represents calendar years. The main geographical clades are highlighted.

640 **Supporting Information**

641 **Supplementary Table 1.** Years, codes, host, localities and country codes of *Zika virus* sequences
642 included in the dataset

643 **Supplementary Table 2.** Zika virus recombinant strains identified using RDP4 package. A total of
644 4 recombinant strains were significantly identified. The recombination is detected using seven
645 recombination detection methods in RDP4 package. These methods include RDP (designed as R),
646 GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMAERA (C), SISCAN (S) and 3SEQ (Q).

647 **Supplementary Fig. 1 Likelihood mapping of the *Zika virus* sequences.** Each dot represents the
648 likelihoods of the three possible unrooted trees for each quartet randomly selected from the data set:
649 the dots near the corners or sides respectively represent tree-like (fully resolved phylogenies in
650 which one tree is clearly better than the others) or network-like phylogenetic signals (three regions
651 in which it is not possible to decide between two topologies). The central area of the map represents
652 a star-like signal (the region in which the star tree is the optimal tree). The numbers indicate the
653 percentage of dots in the centre of the triangle.

654 **Supplementary Table 3.** tMRCA and locations of the minor clades in the discrete
655 phylogeographical analysis.

656 **Supplementary Video 1.** Animated visualization of the continuous pattern of *Zika virus* dispersion
657 in the Americas in the years 2013-2016.

658 **Supplementary Fig. 2 Population dynamics analysis of ZIKV: Bayesian skyline plot (BSP).**
659 The effective number of infections is indicated on the Y axis, and time on the X-axis. The coloured
660 area corresponds to the credibility interval based on the 95% highest posterior density interval
661 (HPD).

662

663

664 REFERENCES

- 665 Ayllon, T., Campos, R.M., Brasil, P., Morone, F.C., Camara, D.C.P., Meira, G.L.S., Tannich, E., Yamamoto, K.A.,
666 Carvalho, M.S., Pedro, R.S., Schmidt-Chanasit, J., Cadar, D., Ferreira, D.F., Honorio, N.A., 2017. Early
667 Evidence for Zika Virus Circulation among *Aedes aegypti* Mosquitoes, Rio de Janeiro, Brazil. *Emerg*
668 *Infect Dis* 23, 1411-1412.
- 669 Baca-Carrasco, D., Velasco-Hernandez, J.X., 2016. Sex, Mosquitoes and Epidemics: An Evaluation of Zika
670 Disease Dynamics. *Bull Math Biol* 78, 2228-2242.
- 671 Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M.A., Alekseyenko, A.V., 2012. Improving the
672 accuracy of demographic and molecular clock model comparison while accommodating phylogenetic
673 uncertainty. *Molecular biology and evolution* 29, 2157-2167.
- 674 Barzon, L., Pacenti, M., Franchin, E., Lavezzo, E., Trevisan, M., Sgarabotto, D., Palu, G., 2016. Infection
675 dynamics in a traveller with persistent shedding of Zika virus RNA in semen for six months after
676 returning from Haiti to Italy, January 2016. *Euro Surveill* 21, 1560-7917.
- 677 Bielejec, F., Baele, G., Vrancken, B., Suchard, M.A., Rambaut, A., Lemey, P., 2016. SpreaD3: Interactive
678 Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Molecular biology and*
679 *evolution* 33, 2167-2169.
- 680 Bloomquist, E.W., Lemey, P., Suchard, M.A., 2010. Three roads diverged? Routes to phylogeographic
681 inference. *Trends Ecol Evol* 25, 626-632.
- 682 Boskova, V., Stadler, T., Magnus, C., 2018. The influence of phylodynamic model specifications on
683 parameter estimates of the Zika virus epidemic. *Virus Evol* 4, vex044.
- 684 Buechler, C.R., Bailey, A.L., Weiler, A.M., Barry, G.L., Breitbach, M.E., Stewart, L.M., Jasinska, A.J., Freimer,
685 N.B., Apetrei, C., Phillips-Conroy, J.E., Jolly, C.J., Rogers, J., Friedrich, T.C., O'Connor, D.H., 2017.
686 Seroprevalence of Zika Virus in Wild African Green Monkeys and Baboons. *mSphere* 2, 00392-00316.
- 687 Calvet, G.A., Filippis, A.M., Mendonca, M.C., Sequeira, P.C., Siqueira, A.M., Veloso, V.G., Nogueira, R.M.,
688 Brasil, P., 2016. First detection of autochthonous Zika virus transmission in a HIV-infected patient in
689 Rio de Janeiro, Brazil. *J Clin Virol* 74, 1-3.

690 Calvez, E., Mousson, L., Vazeille, M., O'Connor, O., Cao-Lormeau, V.M., Mathieu-Daude, F., Pocquet, N.,
691 Failloux, A.B., Dupont-Rouzeyrol, M., 2018. Zika virus outbreak in the Pacific: Vector competence of
692 regional vectors. *PLoS Negl Trop Dis* 12, e0006637.

693 Campos, G.S., Bandeira, A.C., Sardi, S.I., 2015. Zika Virus Outbreak, Bahia, Brazil. *Emerg Infect Dis* 21, 1885-
694 1886.

695 Chiu, C.Y., Sanchez-San Martin, C., Bouquet, J., Li, T., Yagi, S., Tamhankar, M., Hodara, V.L., Parodi, L.M.,
696 Somasekar, S., Yu, G., Giavedoni, L.D., Tardif, S., Patterson, J., 2017. Experimental Zika Virus
697 Inoculation in a New World Monkey Model Reproduces Key Features of the Human Infection. *Sci Rep*
698 7, 17126.

699 D'Ortenzio, E., Matheron, S., Yazdanpanah, Y., de Lamballerie, X., Hubert, B., Piorkowski, G., Maquart, M.,
700 Descamps, D., Damond, F., Leparac-Goffart, I., 2016. Evidence of Sexual Transmission of Zika Virus. *N*
701 *Engl J Med* 374, 2195-2198.

702 De Smet, B., Van den Bossche, D., van de Werve, C., Mairesse, J., Schmidt-Chanasit, J., Michiels, J., Arien,
703 K.K., Van Esbroeck, M., Cnops, L., 2016. Confirmed Zika virus infection in a Belgian traveler returning
704 from Guatemala, and the diagnostic challenges of imported cases into Europe. *J Clin Virol* 80, 8-11.

705 Djikeng, A., Halpin, R., Kuzmickas, R., Depasse, J., Feldblyum, J., Sengamalay, N., Afonso, C., Zhang, X.,
706 Anderson, N.G., Ghedin, E., Spiro, D.J., 2008. Viral genome sequencing by random priming methods.
707 *BMC Genomics* 9, 5.

708 Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past
709 population dynamics from molecular sequences. *Molecular biology and evolution* 22, 1185-1192.

710 Duffy, M.R., Chen, T.H., Hancock, W.T., Powers, A.M., Kool, J.L., Lanciotti, R.S., Pretrick, M., Marfel, M.,
711 Holzbauer, S., Dubray, C., Guillaumot, L., Griggs, A., Bel, M., Lambert, A.J., Laven, J., Kosoy, O.,
712 Panella, A., Biggerstaff, B.J., Fischer, M., Hayes, E.B., 2009. Zika virus outbreak on Yap Island,
713 Federated States of Micronesia. *N Engl J Med* 360, 2536-2543.

714 Faria, N.R., Quick, J., Claro, I.M., Theze, J., de Jesus, J.G., Giovanetti, M., Kraemer, M.U.G., Hill, S.C., Black,
715 A., da Costa, A.C., Franco, L.C., Silva, S.P., Wu, C.H., Raghwani, J., Cauchemez, S., du Plessis, L.,

716 Verotti, M.P., de Oliveira, W.K., Carmo, E.H., Coelho, G.E., Santelli, A., Vinhal, L.C., Henriques, C.M.,
 717 Simpson, J.T., Loose, M., Andersen, K.G., Grubaugh, N.D., Somasekar, S., Chiu, C.Y., Munoz-Medina,
 718 J.E., Gonzalez-Bonilla, C.R., Arias, C.F., Lewis-Ximenez, L.L., Baylis, S.A., Chieppe, A.O., Aguiar, S.F.,
 719 Fernandes, C.A., Lemos, P.S., Nascimento, B.L.S., Monteiro, H.A.O., Siqueira, I.C., de Queiroz, M.G.,
 720 de Souza, T.R., Bezerra, J.F., Lemos, M.R., Pereira, G.F., Loudal, D., Moura, L.C., Dhalia, R., Franca,
 721 R.F., Magalhaes, T., Marques, E.T., Jr., Jaenisch, T., Wallau, G.L., de Lima, M.C., Nascimento, V., de
 722 Cerqueira, E.M., de Lima, M.M., Mascarenhas, D.L., Neto, J.P.M., Levin, A.S., Tozetto-Mendoza, T.R.,
 723 Fonseca, S.N., Mendes-Correa, M.C., Milagres, F.P., Segurado, A., Holmes, E.C., Rambaut, A., Bedford,
 724 T., Nunes, M.R.T., Sabino, E.C., Alcantara, L.C.J., Loman, N.J., Pybus, O.G., 2017. Establishment and
 725 cryptic transmission of Zika virus in Brazil and the Americas. *Nature* 546, 406-410.

726 Faye, O., Freire, C.C., Iamarino, A., Faye, O., de Oliveira, J.V., Diallo, M., Zanotto, P.M., Sall, A.A., 2014.
 727 Molecular evolution of Zika virus during its emergence in the 20(th) century. *PLoS Negl Trop Dis* 8,
 728 e2636.

729 Gao, D., Lou, Y., He, D., Porco, T.C., Kuang, Y., Chowell, G., Ruan, S., 2016. Prevention and Control of Zika as
 730 a Mosquito-Borne and Sexually Transmitted Disease: A Mathematical Modeling Analysis. *Sci Rep* 6,
 731 28070.

732 Garcia-Luna, S.M., Weger-Lucarelli, J., Ruckert, C., Murrieta, R.A., Young, M.C., Byas, A.D., Fauver, J.R.,
 733 Perera, R., Flores-Suarez, A.E., Ponce-Garcia, G., Rodriguez, A.D., Ebel, G.D., Black, W.C.t., 2018.
 734 Variation in competence for ZIKV transmission by *Aedes aegypti* and *Aedes albopictus* in Mexico.
 735 *PLoS Negl Trop Dis* 12, e0006599.

736 Giovanetti, M., Milano, T., Alcantara, L.C., Carcangiu, L., Cella, E., Lai, A., Lo Presti, A., Pascarella, S.,
 737 Zehender, G., Angeletti, S., Ciccozzi, M., 2016. Zika Virus spreading in South America: Evolutionary
 738 analysis of emerging neutralizing resistant Phe279Ser strains. *Asian Pac J Trop Med* 9, 445-452.

739 Griffiths, R.C., Tavaré, S., 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R*
 740 *Soc Lond B Biol Sci* 344, 403-410.

741 Gubler, D.J., Vasilakis, N., Musso, D., 2017. History and Emergence of Zika Virus. *J Infect Dis* 216, S860-S867.

742 Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and
743 methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst*
744 *Biol* 59, 307-321.

745 Han, J.F., Jiang, T., Ye, Q., Li, X.F., Liu, Z.Y., Qin, C.F., 2016. Homologous recombination of Zika viruses in the
746 Americas. *J Infect* 73, 87-88.

747 Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., Gibson, T.J., 1998. Multiple sequence alignment
748 with Clustal X. *Trends Biochem Sci* 23, 403-405.

749 Kass, R.E., Raftery, A.E., 1995. Bayes factors. *Journal of the american statistical association* 90, 773-795.

750 Lana, R.M., Carneiro, T.G., Honorio, N.A., Codeco, C.T., 2014. Seasonal and nonseasonal dynamics of *Aedes*
751 *aegypti* in Rio de Janeiro, Brazil: fitting mathematical models to trap data. *Acta Trop* 129, 25-32.

752 Lazear, H.M., Diamond, M.S., 2016. Zika Virus: New Clinical Syndromes and Its Emergence in the Western
753 Hemisphere. *J Virol* 90, 4864-4875.

754 Lednicky, J., Beau De Rochars, V.M., El Badry, M., Loeb, J., Telisma, T., Chavannes, S., Anilis, G., Cella, E.,
755 Ciccozzi, M., Rashid, M., Okech, B., Salemi, M., Morris, J.G., Jr., 2016. Zika Virus Outbreak in Haiti in
756 2014: Molecular and Clinical Data. *PLoS Negl Trop Dis* 10, e0004687.

757 Lemey, P., Rambaut, A., Drummond, A.J., Suchard, M.A., 2009. Bayesian phylogeography finds its roots.
758 *PLoS Comput Biol* 5, e1000520.

759 Lemey, P., Rambaut, A., Welch, J.J., Suchard, M.A., 2010. Phylogeography takes a relaxed random walk in
760 continuous space and time. *Molecular biology and evolution* 27, 1877-1885.

761 Liang, D., Leung, R.K.K., Lee, S.S., Kam, K.M., 2017. Insights into intercontinental spread of Zika virus. *PLoS*
762 *One* 12, e0176710.

763 Marchette, N.J., Garcia, R., Rudnick, A., 1969. Isolation of Zika virus from *Aedes aegypti* mosquitoes in
764 Malaysia. *Am J Trop Med Hyg* 18, 411-415.

765 Metsky, H.C., Matranga, C.B., Wohl, S., Schaffner, S.F., Freije, C.A., Winnicki, S.M., West, K., Qu, J., Baniecki,
766 M.L., Gladden-Young, A., Lin, A.E., Tomkins-Tinch, C.H., Ye, S.H., Park, D.J., Luo, C.Y., Barnes, K.G.,
767 Shah, R.R., Chak, B., Barbosa-Lima, G., Delatorre, E., Vieira, Y.R., Paul, L.M., Tan, A.L., Barcellona,

768 C.M., Porcelli, M.C., Vasquez, C., Cannons, A.C., Cone, M.R., Hogan, K.N., Kopp, E.W., Anzinger, J.J.,
 769 Garcia, K.F., Parham, L.A., Ramirez, R.M.G., Montoya, M.C.M., Rojas, D.P., Brown, C.M., Hennigan, S.,
 770 Sabina, B., Scotland, S., Gangavarapu, K., Grubaugh, N.D., Oliveira, G., Robles-Sikisaka, R., Rambaut,
 771 A., Gehrke, L., Smole, S., Halloran, M.E., Villar, L., Mattar, S., Lorenzana, I., Cerbino-Neto, J., Valim, C.,
 772 Degrave, W., Bozza, P.T., Gnirke, A., Andersen, K.G., Isern, S., Michael, S.F., Bozza, F.A., Souza, T.M.L.,
 773 Bosch, I., Yozwiak, N.L., MacInnis, B.L., Sabeti, P.C., 2017. Zika virus evolution and spread in the
 774 Americas. *Nature* 546, 411-415.

775 Mittal, R., Nguyen, D., Debs, L.H., Patel, A.P., Liu, G., Jhaveri, V.M., Si, S.K., Mittal, J., Bandstra, E.S., Younis,
 776 R.T., Chapagain, P., Jayaweera, D.T., Liu, X.Z., 2017. Zika Virus: An Emerging Global Health Threat.
 777 *Front Cell Infect Microbiol* 7, 486.

778 Moreira-Soto, A., Carneiro, I.O., Fischer, C., Feldmann, M., Kummerer, B.M., Silva, N.S., Santos, U.G., Souza,
 779 B., Liborio, F.A., Valenca-Montenegro, M.M., Laroque, P.O., da Fontoura, F.R., Oliveira, A.V.D.,
 780 Drosten, C., de Lamballerie, X., Franke, C.R., Drexler, J.F., 2018. Limited Evidence for Infection of
 781 Urban and Peri-urban Nonhuman Primates with Zika and Chikungunya Viruses in Brazil. *mSphere* 3,
 782 00523-00517.

783 Musso, D., Bossin, H., Mallet, H.P., Besnard, M., Broult, J., Baudouin, L., Levi, J.E., Sabino, E.C., Ghawche, F.,
 784 Lanteri, M.C., Baud, D., 2018. Zika virus in French Polynesia 2013-14: anatomy of a completed
 785 outbreak. *Lancet Infect Dis* 18, e172-e182.

786 Olawoyin, O., Kribs, C., 2018. Effects of multiple transmission pathways on Zika dynamics. *Infect Dis Model*
 787 3, 331-344.

788 Passos, S.R.L., Borges Dos Santos, M.A., Cerbino-Neto, J., Buonora, S.N., Souza, T.M.L., de Oliveira, R.V.C.,
 789 Vizzoni, A., Barbosa-Lima, G., Vieira, Y.R., Silva de Lima, M., Hokerberg, Y.H.M., 2017. Detection of
 790 Zika Virus in April 2013 Patient Samples, Rio de Janeiro, Brazil. *Emerg Infect Dis* 23, 2120-2121.

791 Pettersson, J.H., Bohlin, J., Dupont-Rouzeyrol, M., Brynildsrud, O.B., Alfsnes, K., Cao-Lormeau, V.M., Gaunt,
 792 M.W., Falconar, A.K., de Lamballerie, X., Eldholm, V., Musso, D., Gould, E.A., 2018. Re-visiting the
 793 evolution, dispersal and epidemiology of Zika virus in Asia. *Emerg Microbes Infect* 7, 79.

794 Posada, D., 2008. jModelTest: phylogenetic model averaging. *Molecular biology and evolution* 25, 1253-
795 1256.

796 Rambaut, A., Lam, T.T., Max Carvalho, L., Pybus, O.G., 2016. Exploring the temporal structure of
797 heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* 2, vew007.

798 Sebastian, U.U., Ricardo, A.V.A., Alvarez, B.C., Cubides, A., Luna, A.F., Arroyo-Parejo, M., Acuna, C.E.,
799 Quintero, A.V., Villareal, O.C., Pinillos, O.S., Vieda, E., Bello, M., Pena, S., Duenas-Castell, C.,
800 Rodriguez, G.M.V., Ranero, J.L.M., Lopez, R.L.M., Olaya, S.G., Vergara, J.C., Tandazo, A., Ospina, J.P.S.,
801 Leyton Soto, I.M., Fowler, R.A., Marshall, J.C., 2017. Zika virus-induced neurological critical illness in
802 Latin America: Severe Guillain-Barre Syndrome and encephalitis. *J Crit Care* 42, 275-281.

803 Simon-Loriere, E., Holmes, E.C., 2011. Why do RNA viruses recombine? *Nat Rev Microbiol* 9, 617-626.

804 Simonin, Y., van Riel, D., Van de Perre, P., Rockx, B., Salinas, S., 2017. Differential virulence between Asian
805 and African lineages of Zika virus. *PLoS Negl Trop Dis* 11, e0005821.

806 Strimmer, K., von Haeseler, A., 1997. Likelihood-mapping: a simple method to visualize phylogenetic
807 content of a sequence alignment. *Proc Natl Acad Sci U S A* 94, 6815-6819.

808 Towers, S., Brauer, F., Castillo-Chavez, C., Falconar, A.K.I., Mubayi, A., Romero-Vivas, C.M.E., 2016. Estimate
809 of the reproduction number of the 2015 Zika virus outbreak in Barranquilla, Colombia, and
810 estimation of the relative role of sexual transmission. *Epidemics* 17, 50-55.

811 Weaver, S.C., Costa, F., Garcia-Blanco, M.A., Ko, A.I., Ribeiro, G.S., Saade, G., Shi, P.Y., Vasilakis, N., 2016.
812 Zika virus: History, emergence, biology, and prospects for control. *Antiviral Res* 130, 69-80.

813 Zanluca, C., Melo, V.C., Mosimann, A.L., Santos, G.I., Santos, C.N., Luz, K., 2015. First report of
814 autochthonous transmission of Zika virus in Brazil. *Mem Inst Oswaldo Cruz* 110, 569-572.

815 Zehender, G., Veo, C., Ebranati, E., Carta, V., Rovida, F., Percivalle, E., Moreno, A., Lelli, D., Calzolari, M.,
816 Lavazza, A., Chiapponi, C., Baioni, L., Capelli, G., Ravagnan, S., Da Rold, G., Lavezzo, E., Palu, G.,
817 Baldanti, F., Barzon, L., Galli, M., 2017. Reconstructing the recent West Nile virus lineage 2 epidemic
818 in Europe and Italy using discrete and continuous phylogeography. *PLoS One* 12, e0179679.

819 Zhang, Q., Sun, K., Chinazzi, M., Pastore, Y.P.A., Dean, N.E., Rojas, D.P., Merler, S., Mistry, D., Poletti, P.,
820 Rossi, L., Bray, M., Halloran, M.E., Longini, I.M., Jr., Vespignani, A., 2017. Spread of Zika virus in the
821 Americas. Proc Natl Acad Sci U S A 114, E4334-E4343.

822

823

Figure 1
[Click here to download high resolution image](#)

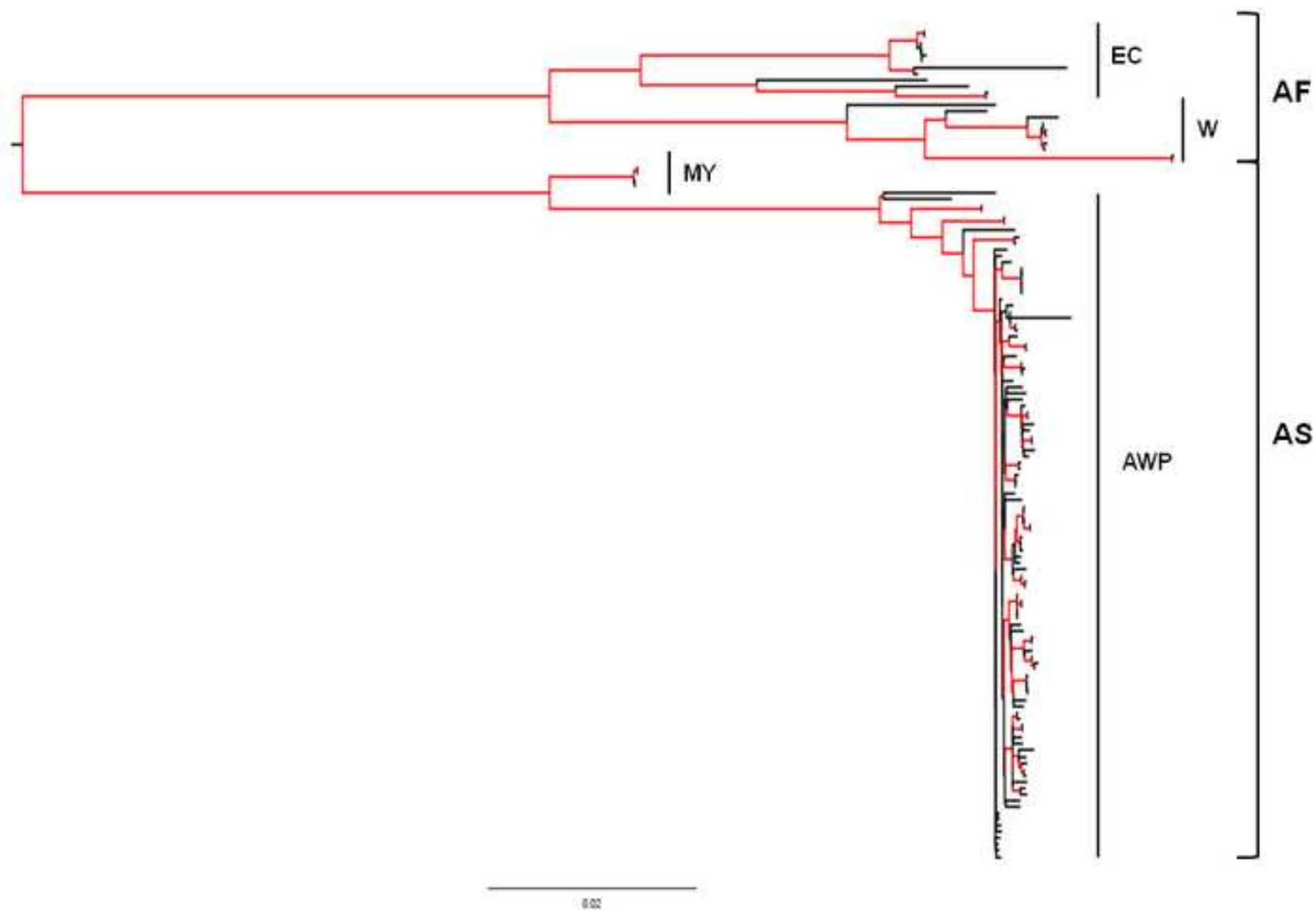


Figure 2
[Click here to download high resolution image](#)

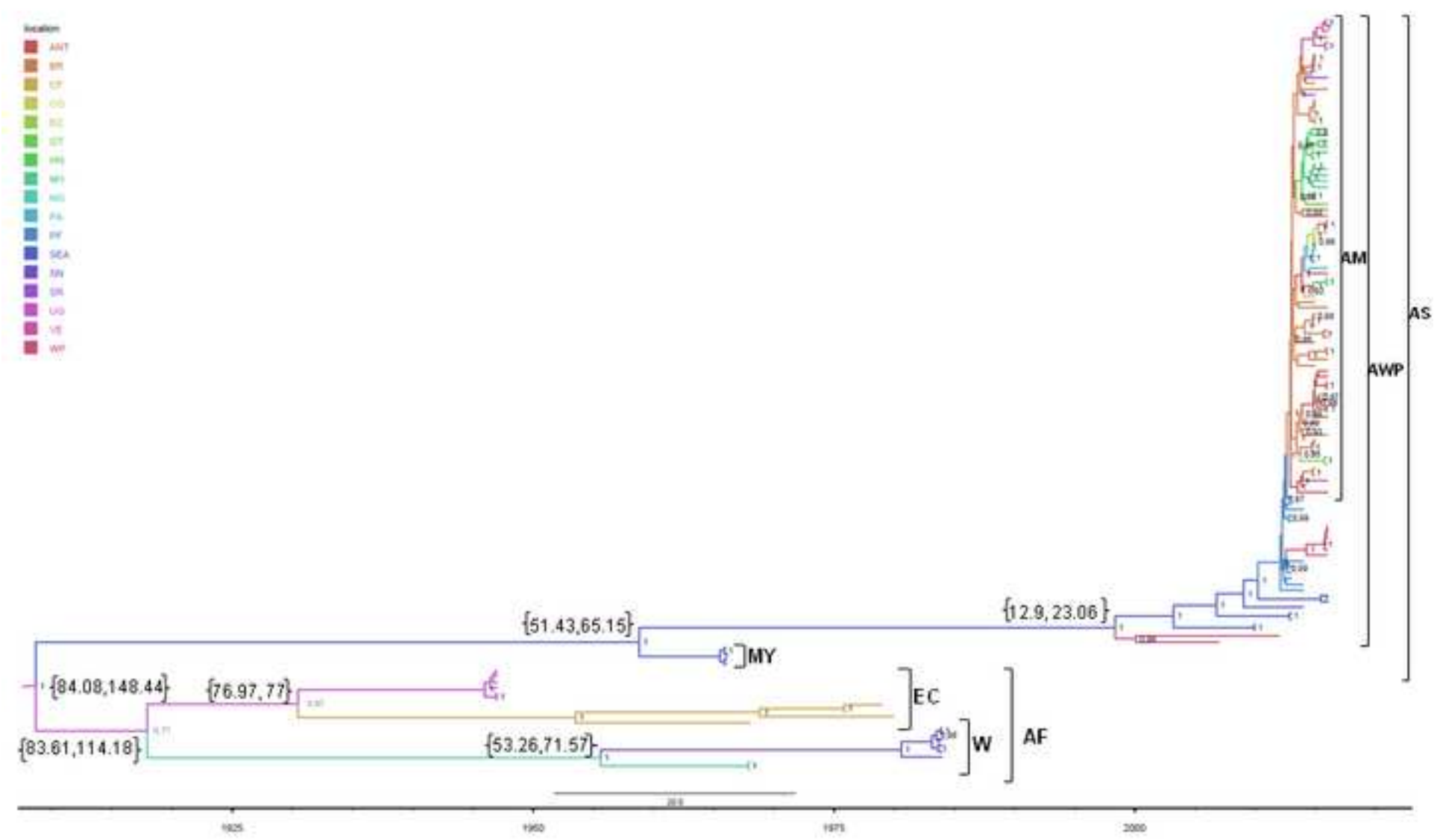


Figure 3
[Click here to download high resolution image](#)

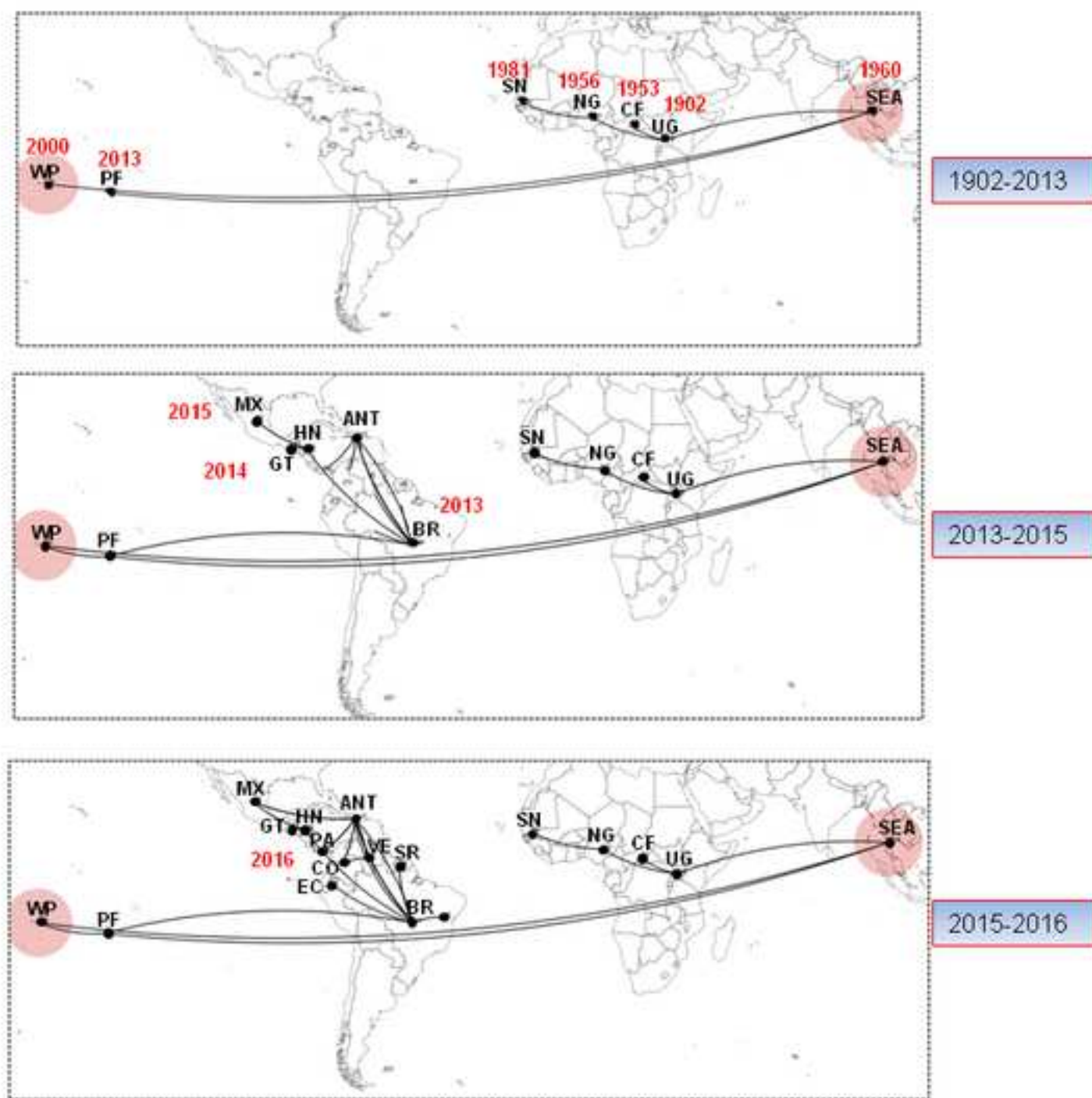


Figure 4
[Click here to download high resolution image](#)

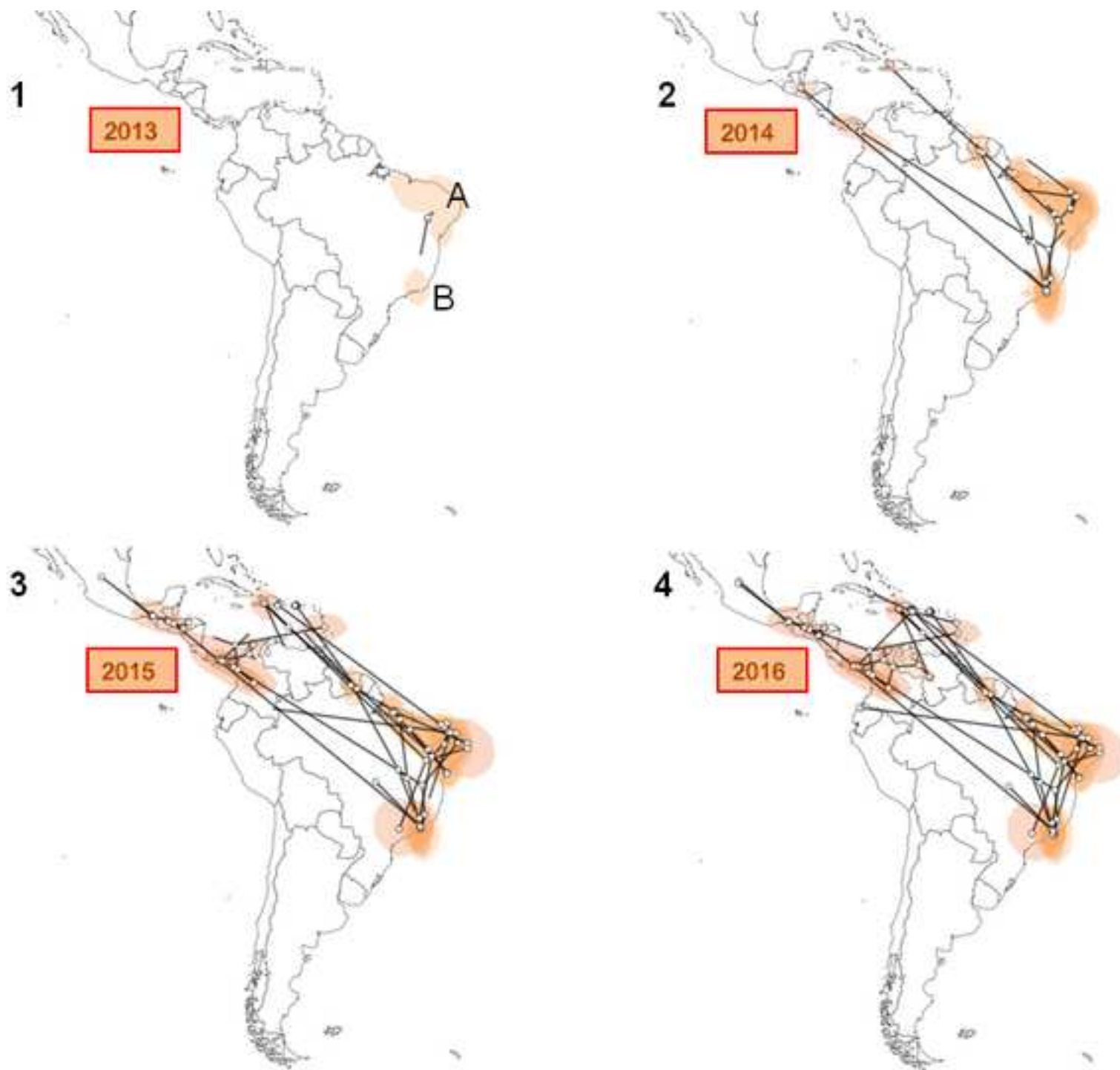
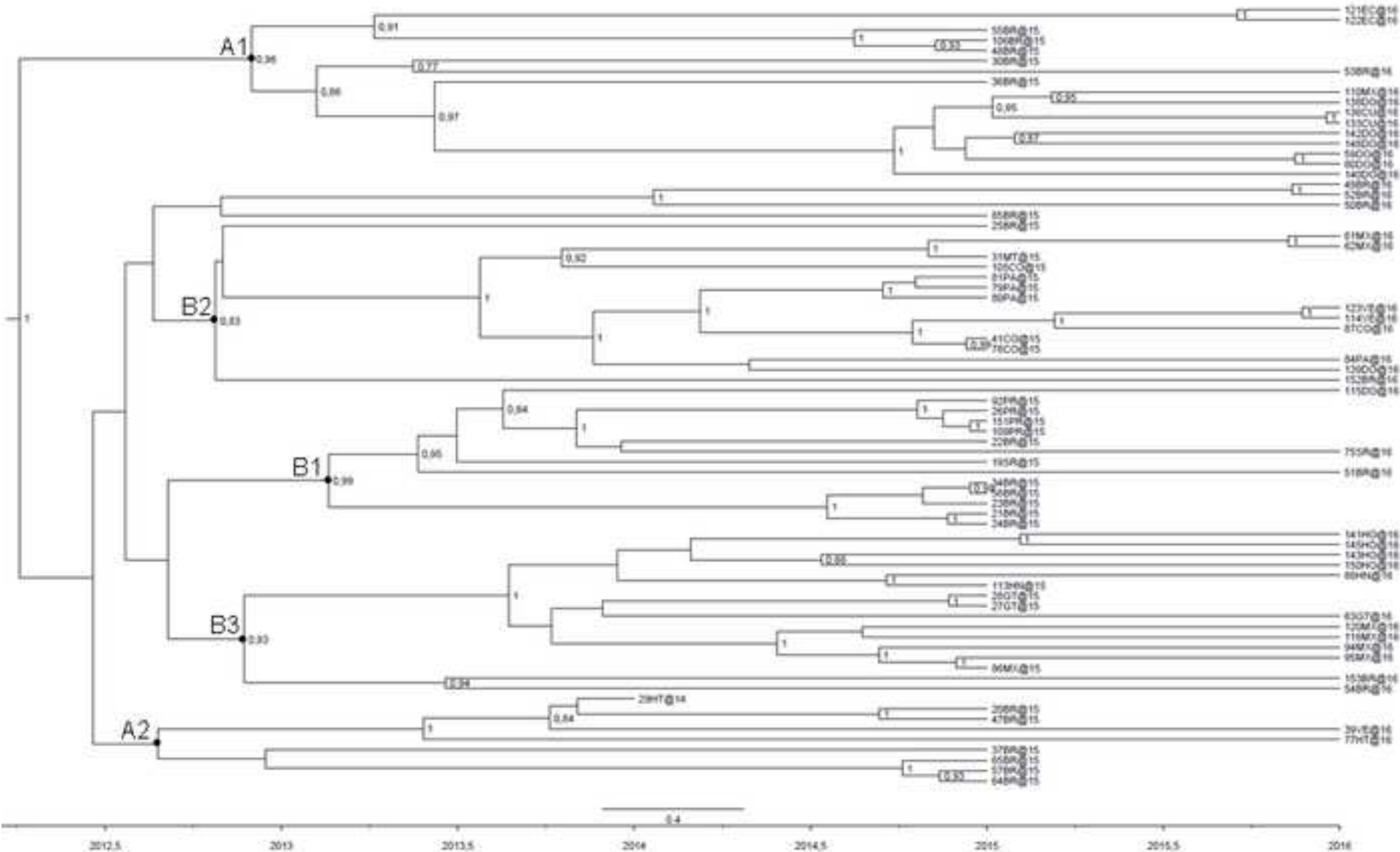


Figure 5
[Click here to download high resolution image](#)



Supplementary Table 1
[Click here to download Supplementary Material: Supplementary Table 1.xlsx](#)

Supplementary Table 2
[Click here to download Supplementary Material: Supplementary Table 2.xlsx](#)

Supplementary Figure 1

[Click here to download Supplementary Material: Supplementary Fig. 1.tif](#)

Supplementary Table 3
[Click here to download Supplementary Material: Supplementary Table 3.xlsx](#)

Supplementary Video 1

[Click here to download Supplementary Material: S1 Video.mp4](#)

Supplementary Figure 2
[Click here to download Supplementary Material: Supplementary Fig. 2.tif](#)

Time-scaled phylogeography of complete *Zika virus* genomes using discrete and continuous space diffusion models

Erika Ebranati^{a,b,1}, Carla Veo^{a,b,1}, Valentina Carta^a, Elena Percivalle^c, Francesca Rovida^c, Elena Rosanna Frati^{b,d}, Antonella Amendola^{b,d}, Massimo Ciccozzi^c, Elisabetta Tanzi^{b,d}, Massimo Galli^{a,b}, Fausto Baldanti^c, Gianguglielmo Zehender^{a,b,*}

^a Department of Biomedical and Clinical Sciences "L. Sacco", University of Milan, Milano, Italy.

^b CRC-Coordinated Research Center "EpiSoMI", University of Milan, Milano, Italy.

^c Molecular Virology Unit, Microbiology and Virology Department, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy.

^d Department of Biomedical Sciences for Health, University of Milan, Milano, Italy.

^e Unit of Medical Statistics and Molecular Epidemiology, University Campus Bio-Medico of Rome, Italy.

*Corresponding author at: Department of Biomedical and Clinical Sciences "L. Sacco", University of Milan, Via G.B. Grassi 74, 20157 Milan, Italy

E-mail address: gianguglielmo.zehender@unimi.it (G. Zehender)

Field Code Changed

¹ These authors contributed equally to this work.

22 ABSTRACT

23 *Zika virus* (ZIKV), a vector-borne infectious agent that has recently been associated with
24 neurological diseases and congenital microcephaly, was first reported in the Western hemisphere in
25 early 2015.

26 A number of authors have reconstructed its epidemiological history using advanced phylogenetic
27 approaches, and the majority of Zika phylogeography studies have used discrete diffusion models.
28 Continuous space diffusion models make it possible to infer the possible origin of the virus in real
29 space by reconstructing its ancestral location on the basis of geographical coordinates deduced from
30 the latitude and longitude of the sampling locations. We analysed all the ZIKV complete genome
31 isolates whose sampling times and localities were available in public databases at the time the study
32 began, using a Bayesian approach for discrete and continuous phylogeographic reconstruction.

33 The discrete phylogeographic analysis suggested that ZIKV emerged to become endemic/epidemic
34 in the first decade of the 1900s in the Ugandan rainforests, and then reached Western Africa and
35 Asia between the 1930s and 1950s. After a long period of about 40 years, it spread to the Pacific
36 islands and reached Brazil from French Polynesia. Continuous phylogeography of the American
37 epidemic showed that the virus entered in north-eastern Brazil in late 2012 and started to spread in
38 early 2013 from two high probability regions: one corresponding to the entire north-east Brazil and
39 the second surrounding the city of Rio de Janeiro, in a mainly northwesterly direction to Central
40 America, the north-western countries of south America and the Caribbean islands. Our data suggest
41 its cryptic circulation in both French Polynesia and Brazil, thus raising questions about the
42 mechanisms underlying its undetected persistence in the absence of a known animal reservoir, and
43 underline the importance of continuous diffusion models in making more reliable phylogeographic
44 reconstructions of emerging viruses.

45

46 **KEYWORDS**

- 47 Zika virus
- 48 Continuous phylogeography
- 49 Surveillance
- 50 Phylodynamics

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76 **1. INTRODUCTION**

77 *Zika virus* (ZIKV) is an emerging arbovirus belonging to the *Flaviviridae* family, genus *Flavivirus*,
78 and is closely phylogenetically related to other important mosquito-borne flaviviruses such as
79 *Japanese encephalitis*, *West Nile* and *Dengue* viruses (Gubler et al., 2017). It was first discovered in
80 a rhesus macaque monkey with fever kept in captivity in the Zika forest on the Entebbe peninsula in
81 Uganda in 1947 (Buechler et al., 2017).

Field Code Changed

Field Code Changed

82 The viral genome is represented by a single-stranded positive-sense RNA molecule of 10.7 kbp that
83 encodes a single polyprotein encompassing three structural proteins (the capsid, the precursor
84 membrane and the envelope proteins) and seven non-structural proteins (NS1, NS2A, NS2B, NS3,
85 NS4A, NS4B and NS5), which play essential roles in virus replication, virulence and secretion
86 (Mittal et al., 2017). Phylogenetic studies of whole ZIKV genomes have revealed the existence of
87 two major evolutionary lineages: one African and the other Asian (Simonin et al., 2017),
88 encompassing also isolates from Pacific Islands and America.

Field Code Changed

Field Code Changed

89 ZIKV is naturally maintained by two distinct transmission cycles: a sylvatic cycle involving non-
90 human primates and arboreal mosquitoes of the genus *Aedes*, and an urban cycle involving humans
91 and urban mosquitoes (mainly *A. aegypti*). It can also be transmitted without vectors: vertically
92 from an infected mother to her child during pregnancy, sexually (Barzon et al., 2016; D'Ortenzio et
93 al., 2016), or by means of blood transfusions or exposure in a laboratory or healthcare setting
94 (Lazear and Diamond, 2016). As with other arboviruses, about 80% of ZIKV-infected subjects are
95 asymptomatic; symptomatic subjects most frequently experience flu-like syndrome, an itchy
96 maculopapular rash and arthritis or arthralgia, but some cases of retro-orbital pain, headache,
97 myalgia and vomiting have also been observed (Giovanetti et al., 2016). ZIKV can cause severe
98 neurological complications such as Guillain-Barré syndrome and microcephaly in infants born to
99 ZIKV-infected women, as demonstrated by the presence of the virus in the brain, placenta or serum
100 of aborted fetuses and newborns with microcephaly (Sebastian et al., 2017). Since the early 1950s,

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

101 ZIKV infection outbreaks have been reported in tropical Africa, south-eastern Asia and the Pacific
 102 islands. The virus was first isolated in Malaysia in 1969 and, later, in Indonesia (Marchette et al.,
 103 1969). In 2007, it caused the first large and well-characterised outbreak on Yap Island, a part of the
 104 Federated States of Micronesia (Duffy et al., 2009), and this was followed by a major epidemic in
 105 French Polynesia in 2013-2014 that affected more than 28,000 people (11% of the population)
 106 (Musso et al., 2018) after which it spread to other neighbouring islands in the south Pacific.
 107 In early 2015, the autochthonous transmission of ZIKV in the northeastern part of Brazil was the
 108 first reported description of the infection in the Americas (Calvet et al., 2016) and, by the end of the
 109 same year, ZIKV activity had expanded into at least 14 Brazilian states (<https://www.paho.org>).
 110 Other indigenous cases of ZIKV infection were detected in Colombia, Suriname, Paraguay and
 111 Venezuela in south America; Guatemala, El Salvador and Mexico in Central America, and
 112 Martinique and Puerto Rico in the Caribbean (Garcia-Luna et al., 2018). In early 2016, local
 113 outbreaks were confirmed in Guyana, Ecuador, Bolivia, Peru, Nicaragua, Curacao, Jamaica, Haiti,
 114 Santo Domingo and other Caribbean islands. In the first half of the same year, the virus was also
 115 detected in Argentina and Cuba and, finally, in the spring 2016, it reached the United States
 116 (Florida) (WHO data available at <http://www.who.int/emergencies/zika-virus/history/en/>). The
 117 epidemics started to decline in various American countries in the second half of 2016 and, although
 118 small local outbreaks were still being reported in 2017, their incidence was greatly reduced.
 119 Ultimately, a total of 48 countries in the Americas had more than 540,000 autochthonous cases
 120 (more than 200,000 in Brazil) and there were about 2,610 reported congenital infections (2,300 in
 121 Brazil) [WHO-PAHO: Regional Zika Epidemiological Update (Americas) August 25, 2017,
 122 available at <https://www.paho.org>].
 123 ZIKV has now become endemic not only in South America and Caribbean, but also in several
 124 Pacific islands (American Samoa, the Federated States of Micronesia, Fiji, Marshall Islands, New
 125 Caledonia, Samoa and Tonga) (Calvez et al., 2018). In addition, there have been an increasing

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

126 number of travel-related cases in non-endemic countries such as Australia, Belgium, Canada, China,
127 France, Portugal, Spain, Switzerland and The Netherlands (De Smet et al., 2016).

128 Several authors have attempted to study the dynamics of Zika virus infection through discrete
129 phylogeographical analysis (Boskova et al., 2018; Faria et al., 2017; Giovanetti et al., 2016; Liang
130 et al., 2017; Metsky et al., 2017; Pettersson et al., 2018).

131 In addition to classical discrete phylogeographical methods, new continuous diffusion models based
132 on Brownian or random walk diffusion models have been developed that can infer ancestral states
133 on the basis of the coordinates of a two-dimensional space identifying the tips of the tree (sampling
134 location). These models allow a more realistic reconstruction of spatial movements because, unlike
135 discrete models, they do not necessarily need the ancestral location to be represented in the
136 sampling location set (Lemey et al., 2010). The differences between discrete and continuous
137 phylogeographic models have been efficiently described in previous reviews (Bloomquist et al.,
138 2010; Faria et al., 2017).

139 The aim of this study was to infer the origin and dispersion routes of ZIKV in the world using a
140 classical discrete method and to reconstruct the recent epidemic in the Americas using a continuous
141 phylogeographical method to better describe the local spread of ZIKV and to make hypothesis
142 about the eco/epidemiology of the virus.

143

144 2. MATERIALS AND METHODS

145 2.1 Patients and datasets

146

147 The study was conducted using 135 complete viral genome sequences retrieved from public
148 databases. These sequences were derived from mosquitoes (n = 21), a sentinel rhesus monkey (n =
149 6), and human samples (n = 108). The sequences were isolated in various countries of the world and

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

retrieved from GenBank (at <http://www.ncbi.nlm.nih.gov/genbank/>); only the sequences with a known location and sampling date were considered.

The sampling period ranged from 1947 to 2016, and the sampling locations ranged from the Central African Republic (CF, n=4) to Nigeria (NG, n=2), Senegal (SN, n=10), Uganda (UG, n=6), south-east Asia (SEA, including Malaysia n=4, Cambodia n=2, Thailand n=3, and Singapore n=2), French Polynesia (FP, n=11), Mexico (MX, n=8), Honduras (HN, n=6), Guatemala (GT, n= 3), Panama (PA, n=4), Venezuela (VE, n=8), Colombia (CO, n=4), Ecuador (EC, n= 2), Brazil (BR, n=28), Suriname (SR, n=2), the Western Pacific (WP, including American Samoa n=6, Tonga n=1, the Philippines n=1, Micronesia n=1), the Antilles (ANT, including Puerto Rico n=4, Haiti n=2, Dominican Republic n=8, Martinique n=1, Cuba n=2). Two of the 28 Brazilian samples were from tourists returning to Italy who became infected in Bahia State.

The sequences were selected on the basis of the following criteria: i) they had to have been published in peer-reviewed journals; ii) their non-recombinant subtype assignment had to be certain; and iii) the city/state of origin and year of sampling had to be known and clearly established in the original publication. The origins and characteristics of the Zika strains dataset are summarised in **Supplementary Table 1**.

2.2 Ethics Statement

Informed consent was obtained according to Italian law (art.13 D.Lgs 196/2003) as well as approval of the Institutional review board of Fondazione IRCCS Policlinico San Matteo on the use of residual biological specimens (IRB Protocol 20100000348).

173 **2.3 Whole genome characterization by means of next-generation** 174 **sequencing** 175

176 The whole ZIKV genome sequence of one human isolate (an Italian semen sample from a traveller
177 coming from Bahia State in Brazil) was previously obtained by Sanger methodology and deposited
178 in GenBank database with the accession number KY003154. Subsequently, in our laboratory, we
179 amplified the whole genome by using the sequence-independent single-primer amplification (SISPA
180 method) and re sequenced it by NGS method (Djikeng et al., 2008).

Field Code Changed

181 ZIKV was isolated on Vero E6 cells; the RNA was prepared by extracting it from the cell culture
182 supernatant and then it was reverse-transcribed using the random primer FR26RV-N
183 (5'GCCGGAGCTCTGCAGATATCNNNNNN3') at a concentration of 10 µM. Viral cDNA was
184 denatured at 94°C for three minutes, and chilled on ice for two minutes. Five units of Klenow
185 fragment (New England Biolabs, Ipswich, MA) were directly added to the reaction to perform the
186 second strand cDNA synthesis. The incubation was carried out at 37°C for one hour, and at 75°C
187 for 10 minutes.

188 Next, 5 µL of double-stranded DNA were added to a PCR master mix containing 5 µL of 10x
189 AccuPrime PCR buffer I, 0.2 µL of AccuPrime Taq DNA Polymerase high fidelity, 4 µL of 10 µM
190 FR20RV (5'GCCGGAGCTCTGCAGATATC3') and 35.8 µL of water. The incubation was
191 performed under the following thermal conditions: 94°C for two minutes, 40 cycles of 94°C for 30
192 seconds, 55°C for one minute and 68°C for three minutes.

193 The PCR product was purified and quantified using a TECAN plate reader. The sample was diluted
194 to an initial concentration of 0.2 ng/µL in accordance with the Illumina protocol, and 1 ng was used
195 for the library preparation (Nextera XT sample preparation Kit, Illumina Inc., San Diego,
196 California, USA).

197 Genomic libraries were sequenced on the Illumina MiSeq platform (Illumina, Inc.) with 2x151 base
198 pairs paired-end runs. Finally, we evaluated the obtained reads for sequence quality and read-pair
199 length using FastQC ver. 0.11.5
200 The reads were assembled using Geneious software v. 11.1.5 (Biomatters , New Zealand) and re-
201 sequencing analysis was performed with the reference virus (KY003154).

202

203 **2.4 Recombination detection**

204

205 In order to identify recombinant strains and exclude them from the analysis, we used the RDP4
206 package, which allows the identification of potential recombinant sequences and their parents
207 (major and minor). It uses seven different methods: RDP (Martin et al., 2015), BOOTSCAN
208 (Martin et al., 2005), CHIMAERA (Posada and Crandall, 2001), SISCAN (Gibbs et al., 2000),
209 GENCONV (Padidam et al., 1999), 3SEQ (Boni et al., 2007) and MAXCHI (Smith, 1992), each of
210 which has a highest acceptable p value of 0.05 and Bonferroni's correction for multiple comparisons
211 The sequences indicated as being recombinant by at least three of these methods were excluded
212 from further analysis.

213 We also screened our alignment using Genetic Algorithm Recombination Detection (GARD)
214 software in order to detect any sequences involved in putative recombinations, and define the
215 number and location of breakpoints (Kosakovsky Pond et al., 2006).

216

217 **2.5 Likelihood mapping**

218

219 The phylogenetic signal of the complete genome dataset was investigated by means of the
220 likelihood mapping (LM) analysis of 10,000 random quartets generated using TreePuzzle (Strimmer
221 and von Haeseler, 1997). Groups of four randomly chosen sequences (quartets) were evaluated and,
222 for each quartet, the three possible unrooted trees were reconstructed using the maximum likelihood

Formatted: Heading 2, Left, Line
spacing: single

Formatted: Font: 16 pt, Font color:
Black

Field Code Changed

223 approach under the selected substitution model that was the General Time Reversible(GTR) with
224 gamma distributed rates among sites. The posterior probabilities of each tree were then plotted on a
225 triangular surface so that fully resolved trees fell into the corners, and the unresolved quartets in the
226 centre of the triangle (a star-tree). When using this strategy, if more than 30% of the dots fall into
227 the centre of the triangle, the data are considered unreliable for the purposes of phylogenetic
228 inference.

229 **2.65 Phylogenetic reconstruction**

230

231 The sequences were aligned using ClustalX software (Jeanmougin et al., 1998) followed by manual
232 editing using Bioedit software v. 7.2.6, and the best fitting nucleotide substitution model was tested
233 by means of the hierarchical likelihood ratio test (LRT) implemented in J Modeltest software
234 (Posada, 2008). The selected model was GTR with gamma distributed rates among sites.

235 The phylogeny of the complete genome was reconstructed using a maximum likelihood approach
236 and the new hill-climbing algorithm implemented in PhyML v.3.0. The reliability of the observed
237 clades was established on the basis of internal node bootstrap values of $\geq 70\%$ (after 200 replicates)
238 (Guindon et al., 2010).

239 The phylogeny of the complete genome was also reconstructed using a Bayesian Markov Chain
240 Monte Carlo (MCMC) method (Beast v. 1.8.4 freely available at <http://beast.bio.ed.ac.uk>). The
241 reliability of the observed clades was established on the basis of posterior probabilities values with
242 significance levels of ≥ 0.7 .

243 The evolutionary rates were estimated under strict and relaxed (with an uncorrelated log normal rate
244 distribution) clock conditions.

245 As coalescent priors, three parametric demographic models of population growth (constant size,
246 exponential growth and logistic growth) the Bayesian SkyGrid, the Bayesian skyline plot (BSP) and
247 the GMRF Bayesian Skyride were compared (Drummond et al., 2005). The best fitting models were
248 selected using the BF implemented in Beast. In accordance with Kass and Raftery, the strength of

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

249 the evidence against H_0 was evaluated as $2\ln BF < 2$ = no evidence; 2-6 = weak evidence; 6-10 =
250 strong evidence, and >10 = very strong evidence. A negative $2\ln BF$ indicates evidence in favour of
251 H_0 . Only values of ≥ 6 were considered significant (Kass and Raftery, 1995).

Field Code Changed

252 We also used path sampling (PS) and stepping stone sampling (SS) to improve the accuracy of
253 model selection (Baele et al., 2012).

Field Code Changed

254 The chains were run for 250 million generations until reaching convergence and sampled every
255 25000 steps. Convergence was assessed by estimating the effective sampling size ($ESS = >200$)
256 after a 10% burn-in, using Tracer software version 1.6 (<http://tree.bio.ed.ac.uk/software/tracer/>). All
257 of the parameters had an ESS of >200 . Uncertainty in the estimates was indicated by 95% highest
258 posterior density (95% HPD) intervals.

Field Code Changed

259 The TMRCA estimates were expressed as the median and 95% HPD years before the most recent
260 sampling date, which corresponded to 2016 in this study.

261

262 **2.6 Recombination detection**

263
264 ~~In order to identify recombinant strains and exclude them from the analysis, we used the RDP4~~
265 ~~package, which allows the identification of potential recombinant sequences and their parents~~
266 ~~(major and minor). It uses seven different methods: RDP (Martin et al., 2015), BOOTSCAN~~
267 ~~(Martin et al., 2005), CHIMAERA (Posada and Crandall, 2001), SISCAN (Gibbs et al., 2000),~~
268 ~~GENCONV (Padidam et al., 1999), 3SEQ (Boni et al., 2007) and MAXCHI (Smith, 1992), each of~~
269 ~~which has a highest acceptable p value of 0.05 and Bonferroni's correction for multiple comparisons~~
270 ~~The sequences indicated as being recombinant by at least three of these methods were excluded~~
271 ~~from further analysis.~~

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

272 ~~We also screened our alignment using Genetic Algorithm Recombination Detection (GARD)~~
273 ~~software in order to detect any sequences involved in putative recombinations, and define the~~
274 ~~number and location of breakpoints (Kosakovsky Pond et al., 2006).~~

Field Code Changed

275

276 **2.7 Root-to-tip regression analysis**

277

278 In order to verify the correlations between time and genetic distances and identify the correct root
279 under the hypothesis of proportionality between them, we used Tempest software and the ML tree
280 to make a root-to-tip regression analysis (Rambaut et al., 2016).

281

282 **2.8 Bayesian phylogeographic analyses**

283

284 **2.8.1 Discrete phylogeographic analysis**

285

286 An improvement to Beast allows an ancestral reconstruction of discrete states in the Bayesian
287 framework described above in which the spatial diffusion of the time-scaled genealogy is modelled
288 as a continuous-time Markov chain process over discrete sampling locations. A Bayesian stochastic
289 search variable selection (BSSVS) approach, which allows the exchange rates in the CTMC to be
290 zero with some prior probability, was used in order to find a minimal (parsimonious) set of rates
291 explaining the diffusions in the phylogeny. Comparing the posterior to prior odds that individual
292 rates are zero provides a Bayes factor test to identify the rates contributing to the migration
293 pathway, which were calculated as described elsewhere. Rates yielding a BF of >3 were considered
294 significant (Lemey et al., 2009).

295 The obtained trees were summarised in a maximum clade credibility tree using the Tree Annotator
296 program included in the Beast package, choosing the tree with the maximum product of posterior
297 probabilities (maximum clade credibility: MCC) after a 10% burn-in. The most probable location of
298 each node was highlighted by labelling the branches with different state colours. In order to
299 visualize diffusion rates over time, it is possible to convert the location-annotated MCC tree to a
300 keyhole mark-up language file (KML) suitable for viewing with georeferencing software.

Field Code Changed

Field Code Changed

301 In order to visualize diffusion rates over time, it is also possible to render the location-annotated
302 MCC tree to a GeoJSON data format suitable for viewing with georeferencing software. The new
303 SPREAD3 analysis tool was used, the MCC tree was converted to a JavaScript object notation
304 (JSON) file and the visualization was rendered using a Data Driven Document (D3) library
305 (Bielejec et al., 2016).

Field Code Changed

306 **2.8.2 Continuous phylogeographic analysis**

307

308 In order to study the spread of ZIKV in more detail, a continuous space phylogeographical analysis
309 was made using American isolates.

310 American ZIKV epidemics were investigated in continuous space using Beast v. 1.8.4. The
311 unknown coordinates were estimated under a strict Brownian diffusion model, and compared with
312 two relaxed random walk (RRW) models relaxing the diffusion rate constancy assumption that
313 respectively assumed the gamma and Cauchy distribution of diffusion rates over the phylogeny
314 (Lemey et al., 2010). Bayes factor comparisons of the models were made by estimating marginal
315 likelihood using path sampling (PS) and stepping stone approaches (Baele et al., 2012). The
316 phylogeny was spatially projected and converted into KML in order to visualize dispersal over time.
317 Uncertainties in the ancestral location estimates were represented by KML polygons delimiting the
318 high-probability regions.

Field Code Changed

Field Code Changed

319 **3. RESULTS**

320

321 **3.1 Illumina paired-end sequencing**

322

323 The raw data reads with quality value $QV > 20$ were filtered by excluding contaminants such as
324 adapters, the ambiguous “N” nucleotides and low-quality sequences using trimming options
325 implemented in Geneious software. After trimming the raw data, a total of 178846 filtered clean
326 reads were obtained with a 76023X coverage.

Field Code Changed

Formatted: Heading 2, Left, Line spacing: single

Formatted: Font: 16 pt, Font color: Black

Formatted: Font: (Default) Calibri, 11 pt

Formatted: Left, Line spacing: Multiple 1.15 li

3.2 Recombination analysis

Both the GARD and RDP programs confirmed the presence of recombination in the final alignment. In particular, RDP analysis (Supplementary Table 2) detected four genomes showing a total of 12 significant recombination events corresponding to isolates from Senegal (12SN@68, 13SN@97, 14SN@01 and 15SN@01). For this reason, these sequences were removed from the definitive data set.

3.32 Likelihood mapping analysis

The presence of phylogenetic noise was investigated using LM analysis. The complete genome data set gave satisfactory results as 7.8% of the dots fell into the central area of the triangles and 87.4% at the corners, thus suggesting that the alignment contained sufficient phylogenetic information (Supplementary Fig. 1).

3.43 Phylogenetic analysis

ML analysis of the whole genomes showed two statistically supported clades (bootstrap=1000) corresponding to the previously described African (AF) and Asian (AS) clades (**Fig. 1**). Two different sub-clades could be distinguished within the African clade: the first (eastern central African, EC) sub-clade included the original 1947 Ugandan isolates, four genomes from the Republic of Central Africa obtained between 1968 and 1980, and two Senegalese isolates obtained in 2001 (each of them grouping together on a geographical basis); the second (Western African, W) sub-clade encompassed the majority of the Senegalese isolates obtained between 1968 and 1997 and two Nigerian strains obtained in 1968, all significantly segregating on the basis of their geographical origin.

The Asian/Pacific Ocean clade (AS) included two geographically distinct monophyletic groups: the first including all of the isolates from Malaysia 1966 (MY), and the second and largest group

(AWP) including all of the other Asian, Western Pacific Ocean, Polynesian and American isolates connected by a long branch (indicating a bottleneck) to the Malaysian clade. The American sub-clade was statistically sustained (bootstrap=1000). Analysis of the single genes highly supported these clades and sub-clades (Fig. 1).

3.4 Recombination analysis

~~Both the GARD and RDP programs confirmed the presence of recombination in the final alignment. In particular, RDP analysis (Supplementary Table 2) detected four genomes showing a total of 12 significant recombination events corresponding to isolates from Senegal (12SN@68, 13SN@97, 14SN@01 and 15SN@01). For this reason, these sequences were removed from the definitive data set.~~

3.5 Root-to-tip regression analysis

Analysis of the unrooted ML tree of the entire dataset without the recombinant sequences, showed a very strong association between genetic distances and sampling dates ($R^2=0.93$, correlation coefficient=0.97), thus confirming the suitability of the dataset for molecular clock analysis. Tempest also located the best tree root within the branch connecting the African and Asian clades. Interestingly, separate analysis of each gene (Table 1) showed similar results in all the datasets, with the weakest temporal signal being obtained using the Membrane dataset ($R^2=0.5$).

Table 1. Comparison between regression analysis and phylogeographic analysis for each single *Zika virus* gene.

ROOT-TO-TIP					BAYESIAN ANALYSIS	
SLOPE (*10 ⁻³)	tMRCA ¹	CORRELATION COEFFICIENT	R SQUARED	RESIDUAL MEAN SQUARED	E.R ² MEAN (95% HPD ³ LOWER-UPPER)	tMRCA ¹

CAPSID	1.19	1878.3	0.97	0.94	2.97*E ⁻⁵	1.03 (0.42-1.6)	1936.5
MEMBRANE	1.02	1885.4	0.73	0.54	3.2*E ⁻⁴	2.2 (1.06-3.5)	1944.8
ENVELOPE	0.85	1802.2	0.91	0.84	4.98*E ⁻⁵	1.7 (0.95-2.4)	1940.6
NS1	0.45	1806.1	0.92	0.84	1.32*E ⁻⁵	1.4 (0.92-2)	1944.3
NS2	1.18	1811.3	0.95	0.9	5.48*E ⁻⁵	1 (0.46-1.6)	1924.9
NS3	0.67	1863.2	0.97	0.94	9.99*E ⁻⁶	0.85 (0.53-1.2)	1924.7
NS4	0.59	1826.9	0.95	0.9	1.33*E ⁻⁵	0.86 (0.58-1.3)	1910.1
NS5	0.63	1835.4	0.97	0.93	1.03*E ⁻⁵	0.93 (0.63-1.3)	1913.6

¹ tMRCA: time of the most Recent Common Ancestor

² E.R: Evolutionary Rate

³ HPD: Highest posterior density, substitutions/site/year (*10⁻³)

3.6 Evolutionary rates, tMRCA estimates and Bayesian phylogeography

The evolutionary rates, tMRCAs and phylogeography were co-estimated using a Bayesian framework implemented in Beast (v. 1.8.4). The comparison by Bayes factor of the marginal likelihoods obtained by applying a strict or relaxed molecular clock under five different coalescent models (2lnBF GMRF Bayesian Skyride vs BSP = - 811.92; 2lnBF constant vs BSP = - 68.31; 2lnBF exponential growth vs BSP = -1092.56; 2lnBF Bayesian Skygrid vs BSP = - 748.1); under a log-normal relaxed clock (2lnBF strict vs relaxed clock = -466.62) showed that the favoured models were the relaxed molecular clock with uncorrelated log-normal rate distribution and the Bayesian skyline plot, the less stringent demographic model. The same model has been confirmed by using path sampling model selection (PS) and stepping stone sampling (SS) model selection (**Table 2**). Under these conditions, we estimated a mean substitution rate for the entire viral genome of 8×10^{-4} (95% HPD $6.4-9.7 \times 10^{-4}$) subs/site/year (**Table 1** shows the mean estimates for each single gene).

The tree-root tMRCA (**Fig. 2**) was estimated to be an average of 114.2 (95% HPD 84-148) years before the present, corresponding to the year 1902 (95% HPD 1868-1932). The MRCA of the eastern Central African clade was placed in 1930 (95% HPD=1918-1940), whereas that of the West

395 African sub-clade originated later, in 1954 (95% HPD=1944-1963). The first Asian node
 396 (corresponding to the Malaysian group) dated back to 1958 (95% HPD 1951-1965), and was
 397 connected by a long branch to the largest AWP sub-clade whose tMRCA dated back to 1998 (95%
 398 HPD 1993-2003). A further highly significant sub-clade (pp=1) included all of the American
 399 strains, which had an estimated mean tMRCA of 3.1 years ago (95% HPD 2.8-3.5), corresponding
 400 to the year 2013 (2012-2014). In general, the isolates of this AWP clade also tended to segregate
 401 significantly on the basis of their geographical origin (**Table 3**).

402 Analysis of migration flows showed 19 (95% HPD=17-21) non-zero rates between different
 403 localities, all of which were significant at BF analysis (BF>3). The suggested dispersion pathway
 404 summarized in **Fig. 3** showed that the currently circulating ZIKV strains shared a common ancestor
 405 that existed in eastern Central Africa in the first decades of the 1900s, and spread to Western Africa
 406 in the 1950s. In the same period, it spread to Asia (Malaysia) and the Asian strain reached the
 407 Pacific islands at least twice: in the first decade of the 2000s and in 2012, when it spread to French
 408 Polynesia. Finally, from French Polynesia, it reached Brazil in 2013 to start a flow that
 409 subsequently reached a several number of central and south American regions.

410 Considering the discrete phylogeographic tree and limiting the analysis to the Asian clade without
 411 the Malaysian strains, we observed a total of 14 highly significant ($0.9 < pp < 1$) monophyletic groups
 412 including more than two isolates (a median of five isolates for clade, range 3-9). The clades were
 413 strongly defined on a geographical basis (**Supplementary Table 3**): five were Brazilian clades
 414 (three including only Brazilian isolates, and two including mixed isolates from Brazil and Ecuador
 415 or Italy); two included Venezuelan isolates in one clade with isolates from Colombia and the other
 416 with one Dominican sequence; two were pure Central American clades (one from Mexico and the
 417 other from Panama) and two others were from the Caribbean (one pure from Puerto Rico and one
 418 mixed from Cuba and Santo Domingo, with one strain from Mexico). The earliest tMRCA was that
 419 of the French Polynesia clade, followed by the Brazilian clade, the Central American clades and,
 420 finally the clades from Antilles and Venezuela.

421 In order to reconstruct the spread of ZIKV in the new world in more detail and avoid the limitations
 422 caused by their arbitrary grouping into discrete localities, we used a continuous phylogeographic
 423 model based on the geographical coordinates of the sampling localities. This analysis only
 424 considered data subset of American isolates. Comparison of the strict Brownian diffusion model
 425 (assuming a homogeneous diffusion rate over the phylogeny) with two RRW models (assuming
 426 different diffusion rates on each branch of the tree) by the BF test showed that a log-normal RRW
 427 diffusion rate fitted the data better than the other models (Cauchy distribution RRW *vs* homogenous
 428 BD: $2\ln BF = 639.34$ by PS and 175.78 by SS; Cauchy distribution RRW *vs* log-normal RRW:
 429 $2\ln BF = 1335.32$ by PS and 10.64 by SS). On the basis of this continuous phylogeographical
 430 reconstruction, the tree root was placed between the coordinates -41.13 E of longitude and -9.53 N
 431 of latitude, corresponding to a location in the state of Bahia in north-east Brazil, close its border
 432 with the two other states of Pernambuco and Piauí (see the animated visualization in
 433 **Supplementary Video 1**). Two high probability (80% HPD) regions (**Fig. 4**) were identified almost
 434 simultaneously at the beginning of the epidemic (2013): the first (A in Fig. 4, panel 1) was a large
 435 ellipse with a major axis of about 1700 km and a minor axis of about 700 km that included the tree-
 436 root and encompassed north-east Brazil (the nine states of Alagoas, Bahia, Ceará, Maranhão,
 437 Paraíba, Pernambuco, Piauí, Rio Grande do Norte, and Sergipe); the second (B in Fig. 4, panel 1)
 438 was a smaller circle with a diameter of about 400 km, surrounding the metropolitan area of Rio de
 439 Janeiro in south-east Brazil. These two areas were the origin of different migration flows,
 440 corresponding to the branches and clades highlighted in the continuous phylogeographic tree (Fig.
 441 4, panels 1-4, and **Fig. 5**). Area A gave rise to two main migration pathways: the first
 442 (corresponding to clade A1 in the tree) initially spread across north-east Brazil and subsequently
 443 reached Santo Domingo (2015), Cuba and Ecuador (2016); the second (corresponding to clade A2,
 444 and not significantly segregated from the root of the tree) first reached the island of Haiti (2014)
 445 and then spread to Venezuela (2016). Area B was the origin of three main migratory flows: the first
 446 (B1) reached Suriname in 2013, and then proceeded towards Porto Rico (2015); the second

(corresponding to clade B2) spread to an uncertainty region encompassing Panama and Colombia in 2013 and subsequently (2015/2016) dispersed east towards the Caribbean (Martinica, Santo Domingo) and Venezuela, and north towards Mexico; and the third (B3) reached Honduras and Guatemala before spreading throughout Central America and reaching Mexico in the last year (2016).

Table 4 shows the geographical coordinates of the localities and tMRCA estimates for each of the main clades (root, A1-2, B1-3). The estimated average tree-root tMRCA was 3.7 years ago, corresponding to October 2012 (95% HPD December 2011 to August 2013), and the estimated average tMRCAs of the main clades were between March (clade A2) and August 2013.

Within a few months, ZIKV had spread to the entire continent. It followed a mainly north-westerly pathway at the beginning of the epidemic but, between late 2014 and the beginning of 2015, went in various directions (along an east/west axis between central America and the Caribbean, and even in a south-easterly direction), thus indicating wider viral dispersion throughout the region. The estimated overall diffusion rate was as much as 760 km/year (between about 600 and 900 km/year).

461

462

463

464

465

466

467

Table 2. Model selection using Path Sampling and Stepping Stone Sampling.

CLOCK	MODELS	COMPLETE GENOME PS ¹	COMPLETE GENOME SS ²
Strict	Constant	-37209,53	-37259,37
Strict	Exponential	-37204,45	-37271,32
Strict	Skygrid	-37208,28	-37265,18
Strict	Skyline	-37173	-37235,26
Strict	Skyride	-37417,47	-37499,97

UCLN ³	Constant	-37148,91	-37218,2
UCLN ³	Exponential	-37153,79	-37225,78
UCLN ³	Skygrid	-37151,84	-37196
UCLN ³	Skyline	-37100,32	-37163,92
UCLN ³	Skyride	-37362,07	-37411,22

¹ Path Sampling

² Stepping stone sampling

³ Uncorrelated log normal

Table 3. Estimated times of the most recent common ancestors (tMRCAs) of the main clades and credibility intervals (95% HPD), with calendar years, most probable locations, and state posterior probabilities (spp) of the 131 complete genomes of *Zika virus*.

Nodes	pp ²	tMRCA ¹			YEARS			location	stpp ⁴
		MEAN	95% HPD ³ lower	95% HPD ³ upper	MEAN	95% HPD ³ Lower	95% HPD ³ upper		
Root	1	114,2	148,44	84,08	1902	1868	1932	UG	0,26
African	0,71	98,6	114,18	83,61	1917,4	1901,8	1932,4	UG	0,34
Eastern Central	0,97	86,1	97,77	76	1929,9	1918,2	1940	UG	0,51
Western	1	61,6	71,57	53,26	1954,3	1944,43	1962,7	NG	0,57
Asian	1	57,7	65,15	51,43	1958,2	1950,85	1964,6	SEA	0,81
American	1	3,1	3,46	2,77	2012,9	2012,5	2013,22	BR	0,83

¹ tMRCA: time of the most Recent Common Ancestor

²pp: posterior probability

³HPD: highest posterior density

⁴stpp: state posterior probability

Table 4. The main clades, calendar months, most probable locations with mean estimated coordinates, and posterior probabilities (pp) of the 76 complete genomes of *Zika virus*.

CLADES	pp ¹	LONGITUDE	LATITUDE	LOCATION	MONTH	MONTH L ⁵	MONTH U ⁶
Root tree	1	41.126	9.532	BRA_NE ²	October 2012	August 2013	December 2011

A1	0.96	38.8	7.204	BRA_NE ²	June 2013	May 2014	August 2012
A2	0.12	42.056	8.41	BRA_NE ²	March 2013	November 2013	May 2012
B2	0.82	42.457	20.003	BRA_SE ³	April 2013	January 2014	August 2012
B1	0.99	45.678	13.439	BRA_C ⁴	August 2013	May 2014	December 2012
B3	0.93	43.377	21.865	BRA_SE ³	May 2013	February 2014	October 2012

¹ Posterior probability

² North-east Brazil

³ South-east Brazil

⁴ Centre Brazil

⁵ Lower

⁶ Upper

4. DISCUSSION

A number of authors have recently attempted to estimate the evolutionary dynamics of ZIKV in the Americas with the aim of reconstructing the most probable origin of the epidemic, the time of its entry into the Americas, and the diffusion pathways that led to its spread across the continent in such a short time. Some of these studies used partial coding sequences (Giovanetti et al., 2016; Liang et al., 2017), and others whole viral genomes (Faria et al., 2017; Metsky et al., 2017). Some authors included in their analyses all of the isolates available in public databases obtained in over 70 years since the first isolates in 1947 (Giovanetti et al., 2016; Liang et al., 2017), whereas others concentrated the study only on American (Boskova et al., 2018; Faria et al., 2017) or Asian isolates (Pettersson et al., 2018).

In this study we reconstructed the spatiotemporal dynamics of ZIKV at a global and a local scale by using, for the first time, two different phylogeographic approaches: a discrete and a continuous diffusion model.

In order to investigate the phylogenetic information contained in partial genes, we made a root-to-tip regression analysis that showed a sufficient temporal structure ($R^2 = 0.93$) in the whole genome dataset, whereas the results of analyses of the individual gene datasets were ambiguous in terms of

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

501 locating the best-fitting root position and estimates of the evolutionary rates and root - tMRCA. For
502 these reasons, the dataset was also analysed for the presence of recombination, which revealed four
503 recombinant genomes isolated in Senegal in different years (1968, 1997 and 2001), carrying a total
504 of 12 recombination breakpoints. Homologous recombinations in ZIKV has been previously
505 described (Faye et al., 2014; Han et al., 2016) also in other flaviviruses (Simon-Loriere and Holmes,
506 2011), and may explain some of the discrepancies in dating the origin of the virus, particularly
507 when this is based on partial genomes.

508 However, although it therefore seems to be essential to use whole genomes and exclude
509 recombinant sequences in order to obtain unbiased inferences, complete genomes are relatively
510 scarce because of the difficulties in performing extensive genome sequencing (Metsky et al., 2017).
511 The low level of viremia frequently requires preparatory cultural enrichment of the virus, and so we
512 have developed a protocol for the whole genome sequencing of ZIKV that was implemented on the
513 Illumina NGS platform. The full genome of an Italian sample of human semen that has been
514 previously characterised using a Sanger-based method was further analysed using our NGS
515 protocol, revealing over 99% identity with KY003154. Our sequence was aligned with 130 other
516 complete ZIKV genomes for which the sampling dates and locations were known that were
517 retrieved from public databases at the time the analysis began.

518 Our findings indicate that the mean evolutionary rate of the viral genome was between 8.5 and
519 22×10^{-4} , similar to the estimates previously obtained by other authors (evolutionary rates ranging
520 from 6 to 13×10^{-4}) (Liang et al., 2017). On the basis of this evolutionary rate, we estimated a tree-
521 root tMRCA (varying from 84 to 148 years), corresponding to the late 19th and early 20th centuries.

522 In agreement with Liang (Liang et al., 2017), who only studied partial genes, our data suggest that
523 the Zika virus emerged more than 100 years ago as an endemic/epidemic infection, probably after a
524 period of sylvatic circulation in the rainforests of the east Africa. The virus then spread to West
525 Africa in the 1950s and, a few years later (1958s), to south-east Asia (Malaysia). After a long period
526 of about 40 years without any new isolation, causing a bottleneck effect (highlighted by the long

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

527 branch of the tree connecting the older Malaysian strains to the new SEA isolates), the virus
528 reappeared in the Pacific islands (Micronesia) in the early 2000s, and caused the first significant
529 outbreak on the Island of Yap in 2007. The virus again left south-east Asia and spread to French
530 Polynesia where it caused the largest outbreak ever recorded before that time. Our estimated
531 tMRCA showed that the virus had been present in French Polynesia at least since the second half of
532 2012 even if the first recordable cases were in late 2013 (between the 41st week of 2013 and the 14th
533 week of 2014). Our data confirm that ZIKV reached French Polynesia from south-east Asia not
534 from the Island of Yap, as suggested by others (Pettersson et al., 2018; Weaver et al., 2016).
535 Finally, it moved from French Polynesia to the Americas (Brazil) and spread throughout the
536 continent within a few years. Interestingly, the skyline analysis showed an exponential increase in
537 the effective number of infections from early 2013 to late 2014, which corresponds with the spread
538 of the virus in the Western hemisphere, according to our reconstruction (**Supplementary Fig. 2**).
539 The added value of this study comes from the use of a continuous diffusion model to reconstruct the
540 spread of ZIKV in the Americas. Discrete phylogeography requires the grouping of isolates into
541 categories that can be based on political-administrative boundaries, or other geographically or
542 epidemiologically homogenous areas/populations. However, the grouping is frequently arbitrary
543 and often lacks precision in reconstructing flow rates between different spatial areas, in particular
544 when there are ambiguities in assigning the isolates to one group or another. Continuous
545 phylogeography allows these limitations to be overcome by identifying isolates on the basis of
546 geographical coordinates corresponding to the latitude and longitude of the sampling location. We
547 deduced these coordinates on the basis of the patients' data available, and the use of diffusion
548 models made the reconstruction of the spread of the epidemic in the Americas more precise by
549 allowing ancestral sequences to reside at any location in a continuous bi-dimensional space
550 whereas, in the case of discrete approaches, there is no way to infer the ancestral location from the
551 tree if it is not present in the sampled location set.

Field Code Changed

Field Code Changed

552 Our continuous phylogeographical reconstruction allowed an estimate of the possible geographical
 553 coordinates of the entry location and an estimated 80% probability region covering the entire north-
 554 east of Brazil, thus suggesting that ZIKV entered Brazil in late 2012. Just two years later, in early
 555 2015, cases of a “dengue-like syndrome”, that probably represented the first cases of the ZIKV
 556 epidemic, were reported in two cities: Natal (in the State of Rio Grande del Norte) and Camaçari (in
 557 the metropolitan area of Bahia) (Campos et al., 2015; Zanluca et al., 2015). Both cities are included
 558 in the 80% probability region of our estimated tree-root even if the analysed samples did not include
 559 any coming from these places which underlines the importance of using these models.
 560 Interestingly, the analysis also identified a second high probability region surrounding Rio de
 561 Janeiro at the very beginning of the epidemic (2013) as a further initial area of viral dissemination
 562 across the Americas. A recent study using a data-driven stochastic and spatial epidemic model
 563 considering the period between April 2013 and June 2014 estimated the arrival of ZIKV in the
 564 Americas in August 2013 in Rio de Janeiro or north-east Brazil, where mosquito density and DENV
 565 transmission is highest (Zhang et al., 2017). This is in line with our spatial reconstruction, although
 566 our tree root tMRCA is several months earlier (October 2012, with an upper estimate of August
 567 2013). Furthermore, a recent molecular survey identified the presence of ZIKV RNA in samples
 568 collected in March-May 2013 in Tijuca (in the metropolitan area of Rio de Janeiro) from patients
 569 with acute febrile syndrome negative for DENV RNA (Passos et al., 2017), and other phylogenetic
 570 studies based on human and entomological samples support the presence of ZIKV in Brazil in late
 571 2012 and early 2013 (Ayllon et al., 2017; Metsky et al., 2017).
 572 Two waves of the ZIKV epidemic in Brazil have been hypothesised: an early wave in north-east
 573 Brazil and a second in south-east Brazil (Zhang et al., 2017). Our continuous phylogeographical
 574 analysis of the American clades identified two clades supporting the hypothesis of a first epidemic
 575 wave starting in north-east Brazil in early 2013 that initially spread locally (A1) and reached Haiti
 576 (A2). Previous authors have suggested an initial wave of ZIKV in Haiti (Lednicky et al., 2016). The
 577 same viral strains were only later exported to Santo Domingo, Cuba and Ecuador.

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

578 The second epidemic wave starting in the region surrounding Rio de Janeiro (B) spread over a
579 larger area and give rise to new dispersal centres in Suriname, Panama/Colombia and
580 Honduras/Guatemala from which the infection further spread to the Caribbean islands and Central
581 America. The more extensive spread of the strains originating in region B may have been due to the
582 larger passenger flows related to the importance of Rio as a Brazilian transportation hub (Zhang et
583 al., 2017).

Field Code Changed

584 Globally, the virus moved north-west from Brazil to the central-north America at an estimated mean
585 diffusion rate of 760.8 km/year (95% HPD 596-913 km/year) between early 2013 and 2016.

586 It has recently been suggested that ZIKV may have been imported into the Americas as a
587 consequence of the Confederations Cup held in Brazil in June 2013. It is probable that, during this
588 event, many athletes and supporters from affected areas (French Polynesia) probably travelled
589 through a large area of eastern Brazil stretching from Fortaleza to Rio de Janeiro. An alternative
590 hypothesis is the simultaneous arrival of the same south-east Asian virus in French Polynesia and
591 Brazil (Zhang et al., 2017), but this does not seem to be confirmed by our analysis, which indicates
592 that ZIKV entered Polynesia before reaching Brazil.

Field Code Changed

593 One question that deserves to be clarified is the cryptic circulation of the virus in Brazil before the
594 first cases of ZIKV infection were reported in 2013-2015; this also seems to have occurred in
595 Polynesia, where the estimated first entry of the virus is at least one year before the first human
596 cases were reported. The frequency of asymptomatic infections and the presence in the same area of
597 viruses causing infections with similar outcomes (such as *Chikungunya* and *Dengue* viruses) may
598 have masked the initial spread of the virus. Moreover, sylvatic viral circulation can be hypothesised,
599 even if there is only limited evidence of the exposure of non-human primates to Zika virus in the
600 new world (Chiu et al., 2017; Moreira-Soto et al., 2018). Other factors such as the density of the
601 vector population and the seasonality nature of vector abundance (Lana et al., 2014) may also affect
602 the duration of silent circulation.

Field Code Changed

Field Code Changed

Field Code Changed

603 In the absence of any known animal reservoirs and a possible enzootic circulation such as in the
604 case of *West Nile* Virus, which is known to have undergone a latent period of circulation before the
605 appearance of the first human cases (Zehender et al., 2017), the possible role of transmission routes
606 other than *Aedes* mosquito bites should be investigated. It has been reported that the infection can
607 be sexually transmitted due to its persistent presence in the semen of affected males, even if they are
608 asymptomatic. Recent studies have shown that the sexual transmission may contribute to increase
609 the final size and the persistence of the epidemic (Gao et al., 2016). In particular, a recent study
610 showed that up to 47% of ZIKV cases may be due to sexual contacts when *Aedes* mosquitoes are
611 also present (Towers et al., 2016). Sexual transmission and possibly migration of recently infected
612 subjects (Baca-Carrasco and Velasco-Hernandez, 2016; Olawoyin and Kribs, 2018) may have
613 initially ~~contributed~~ to the slow ~~er~~ and hidden circulation of the virus before the accumulation of
614 a critical number of infected humans and mosquitoes (~~Baca Carrasco and Velasco Hernandez, 2016;~~
615 ~~Olawoyin and Kribs, 2018)~~ transformed the spatially and temporally limited outbreak into an
616 ~~explosively widespread epidemic~~.

617 This is the first study, to our knowledge, that provides an estimate of the geographic origin and
618 diffusion pathways of ZIKV in America across a continuous space. Moreover, it provides a new
619 ZIKV genome obtained through NGS platform.

620 The main limitation of this study is that it analysed a relatively small number of complete ZIKV
621 genomes, although it must be remembered that the databases included only 135 complete genomes
622 at the time the study was started. It is not a limitation in terms of coalescent theory, which considers
623 small samples of whole populations (Griffiths and Tavaré, 1994) but, as it may be a problem for
624 phylogeographical studies in terms of sampled locations, we partially compensated this by using
625 continuous phylogeography. Unfortunately, the scarcity of publicly available sequences is
626 essentially due to the difficulty in detecting the virus in samples and the need for cultures in order to
627 have sufficient material (Metsky et al., 2017)

Field Code Changed

Field Code Changed

Formatted: Font: Italic

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

628 Our data underline the importance of genome characterisation and the use of phylogeography in the
629 surveillance of emerging infections.

630 **Acknowledgements**

631
632 This work was partially financed by the “NANOMAX” Bandiera project

633 (2013–2015) funded by Italian Ministry for Education, University and Research (grant number
634 G42I12000180005) to GZ.

635 This study was also partially supported by funds from Lombardy Region and grant from
636 the Ministero della Salute, Ricerca Corrente Fondazione Istituto Di Ricovero e Cura a
637 Carattere Scientifico Policlinico San Matteo, grant no. 80206 to EP.

638

Figures legend

Fig. 1. Maximum likelihood tree of the 133 Zika virus complete genome sequences. The significant posterior probabilities ($pp \geq 0.7$) of the corresponding nodes have been coloured in red and the main clades have been highlighted. The scale bar indicates 2% of nucleotide divergence.

Fig. 2. Phylogeographic analysis of 131 Zika virus isolates in the world. The branches of the maximum clade credibility (MCC) tree are coloured on the basis of the most probable location of the descendent nodes (ANT= Antille; BR= Brazil; CF= Central African Republic; CO= Colombia; EC=Ecuador; GT= Guatemala; HN= Honduras; MX= Mexico; NG= Nigeria; PA= Panama; PF= French Polynesia; SEA= South Eastern Asia; SN= Senegal; SR= Suriname; UG= Uganda; VE= Venezuela; WP= Western Pacific). The numbers on the internal nodes indicate posterior probabilities \geq to 0.7, and the scale at the bottom of the tree represents calendar years. The main geographical clades are highlighted.

Fig. 3. Spatio-temporal dynamics of the Zika virus epidemic in the world. The figure summarises the most significant migration links in the involved areas.

Fig. 4. Spatio-temporal dynamics of the Zika virus epidemic in the Americas. The figure summarises the most significant migration links in the involved area.

Fig. 5. Phylogeographical analysis of 76 Zika virus isolates in the Americas. The numbers on the internal nodes indicate posterior probabilities of >0.7 , and the scale at the bottom of the tree represents calendar years. The main geographical clades are highlighted.

663 **Supporting Information**

664 **Supplementary Table 1.** Years, codes, host, localities and country codes of *Zika virus* sequences
665 included in the dataset

666 Supplementary Table 2. *Zika virus* recombinant strains identified using RDP4 package. A total of 4
667 recombinant strains were significantly identified. The recombination is detected using seven
668 recombination detection methods in RDP4 package. These methods include RDP (designed as R),
669 GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMAERA (C), SISCAN (S) and 3SEQ (Q).

670 **Supplementary Fig. 1 Likelihood mapping of the *Zika virus* sequences.** Each dot represents the
671 likelihoods of the three possible unrooted trees for each quartet randomly selected from the data set:
672 the dots near the corners or sides respectively represent tree-like (fully resolved phylogenies in
673 which one tree is clearly better than the others) or network-like phylogenetic signals (three regions
674 in which it is not possible to decide between two topologies). The central area of the map represents
675 a star-like signal (the region in which the star tree is the optimal tree). The numbers indicate the
676 percentage of dots in the centre of the triangle.

677 ~~Supplementary Table 2. *Zika virus* recombinant strains identified using RDP4 package. A total of~~
678 ~~4 recombinant strains were significantly identified. The recombination is detected using seven~~
679 ~~recombination detection methods in RDP4 package. These methods include RDP (designed as R),~~
680 ~~GENCONV (G), BOOTSCAN (B), MAXCHI (M), CHIMAERA (C), SISCAN (S) and 3SEQ (Q).~~

681 **Supplementary Table 3.** tMRCA and locations of the minor clades in the discrete
682 phylogeographical analysis.

683 **Supplementary Video 1.** Animated visualization of the continuous pattern of *Zika virus* dispersion
684 in the Americas in the years 2013-2016.

685 **Supplementary Fig. 2 Population dynamics analysis of ZIKV: Bayesian skyline plot (BSP).**
686 The effective number of infections is indicated on the Y axis, and time on the X-axis. The coloured

687 area corresponds to the credibility interval based on the 95% highest posterior density interval
688 (HPD).

689

690

691 **REFERENCES**

692 Ayllon, T., Campos, R.M., Brasil, P., Morone, F.C., Camara, D.C.P., Meira, G.L.S., Tannich, E., Yamamoto, K.A.,
693 Carvalho, M.S., Pedro, R.S., Schmidt-Chanasit, J., Cadar, D., Ferreira, D.F., Honorio, N.A., 2017. Early
694 Evidence for Zika Virus Circulation among *Aedes aegypti* Mosquitoes, Rio de Janeiro, Brazil. *Emerg*
695 *Infect Dis* 23, 1411-1412.

696 Baca-Carrasco, D., Velasco-Hernandez, J.X., 2016. Sex, Mosquitoes and Epidemics: An Evaluation of Zika
697 Disease Dynamics. *Bull Math Biol* 78, 2228-2242.

698 Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M.A., Alekseyenko, A.V., 2012. Improving the
699 accuracy of demographic and molecular clock model comparison while accommodating phylogenetic
700 uncertainty. *Molecular biology and evolution* 29, 2157-2167.

701 Barzon, L., Pacenti, M., Franchin, E., Lavezzo, E., Trevisan, M., Sgarabotto, D., Palu, G., 2016. Infection
702 dynamics in a traveller with persistent shedding of Zika virus RNA in semen for six months after
703 returning from Haiti to Italy, January 2016. *Euro Surveill* 21, 1560-7917.

704 Bielejec, F., Baele, G., Vrancken, B., Suchard, M.A., Rambaut, A., Lemey, P., 2016. Spred3: Interactive
705 Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Molecular biology and*
706 *evolution* 33, 2167-2169.

707 Bloomquist, E.W., Lemey, P., Suchard, M.A., 2010. Three roads diverged? Routes to phylogeographic
708 inference. *Trends Ecol Evol* 25, 626-632.

709 Boskova, V., Stadler, T., Magnus, C., 2018. The influence of phylodynamic model specifications on
710 parameter estimates of the Zika virus epidemic. *Virus Evol* 4, vex044.

711 Buechler, C.R., Bailey, A.L., Weiler, A.M., Barry, G.L., Breitbach, M.E., Stewart, L.M., Jasinska, A.J., Freimer,
 712 N.B., Apetrei, C., Phillips-Conroy, J.E., Jolly, C.J., Rogers, J., Friedrich, T.C., O'Connor, D.H., 2017.
 713 Seroprevalence of Zika Virus in Wild African Green Monkeys and Baboons. *mSphere* 2, 00392-00316.
 714 Calvet, G.A., Filippis, A.M., Mendonca, M.C., Sequeira, P.C., Siqueira, A.M., Veloso, V.G., Nogueira, R.M.,
 715 Brasil, P., 2016. First detection of autochthonous Zika virus transmission in a HIV-infected patient in
 716 Rio de Janeiro, Brazil. *J Clin Virol* 74, 1-3.
 717 Calvez, E., Mousson, L., Vazeille, M., O'Connor, O., Cao-Lormeau, V.M., Mathieu-Daude, F., Pocquet, N.,
 718 Failloux, A.B., Dupont-Rouzeyrol, M., 2018. Zika virus outbreak in the Pacific: Vector competence of
 719 regional vectors. *PLoS Negl Trop Dis* 12, e0006637.
 720 Campos, G.S., Bandeira, A.C., Sardi, S.I., 2015. Zika Virus Outbreak, Bahia, Brazil. *Emerg Infect Dis* 21, 1885-
 721 1886.
 722 Chiu, C.Y., Sanchez-San Martin, C., Bouquet, J., Li, T., Yagi, S., Tamhankar, M., Hodara, V.L., Parodi, L.M.,
 723 Somasekar, S., Yu, G., Giavedoni, L.D., Tardif, S., Patterson, J., 2017. Experimental Zika Virus
 724 Inoculation in a New World Monkey Model Reproduces Key Features of the Human Infection. *Sci Rep*
 725 7, 17126.
 726 D'Ortenzio, E., Matheron, S., Yazdanpanah, Y., de Lamballerie, X., Hubert, B., Piorkowski, G., Maquart, M.,
 727 Descamps, D., Damond, F., Leparac-Goffart, I., 2016. Evidence of Sexual Transmission of Zika Virus. *N*
 728 *Engl J Med* 374, 2195-2198.
 729 De Smet, B., Van den Bossche, D., van de Werve, C., Mairesse, J., Schmidt-Chanasit, J., Michiels, J., Arien,
 730 K.K., Van Esbroeck, M., Cnops, L., 2016. Confirmed Zika virus infection in a Belgian traveler returning
 731 from Guatemala, and the diagnostic challenges of imported cases into Europe. *J Clin Virol* 80, 8-11.
 732 Djikeng, A., Halpin, R., Kuzmickas, R., Depasse, J., Feldblyum, J., Sengamalay, N., Afonso, C., Zhang, X.,
 733 Anderson, N.G., Ghedin, E., Spiro, D.J., 2008. Viral genome sequencing by random priming methods.
 734 *BMC Genomics* 9, 5.
 735 Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past
 736 population dynamics from molecular sequences. *Molecular biology and evolution* 22, 1185-1192.

737 Duffy, M.R., Chen, T.H., Hancock, W.T., Powers, A.M., Kool, J.L., Lanciotti, R.S., Pretrick, M., Marfel, M.,
 738 Holzbauer, S., Dubray, C., Guillaumot, L., Griggs, A., Bel, M., Lambert, A.J., Laven, J., Kosoy, O.,
 739 Panella, A., Biggerstaff, B.J., Fischer, M., Hayes, E.B., 2009. Zika virus outbreak on Yap Island,
 740 Federated States of Micronesia. *N Engl J Med* 360, 2536-2543.

741 Faria, N.R., Quick, J., Claro, I.M., Theze, J., de Jesus, J.G., Giovanetti, M., Kraemer, M.U.G., Hill, S.C., Black,
 742 A., da Costa, A.C., Franco, L.C., Silva, S.P., Wu, C.H., Raghwan, J., Cauchemez, S., du Plessis, L.,
 743 Verotti, M.P., de Oliveira, W.K., Carmo, E.H., Coelho, G.E., Santelli, A., Vinhal, L.C., Henriques, C.M.,
 744 Simpson, J.T., Loose, M., Andersen, K.G., Grubaugh, N.D., Somasekar, S., Chiu, C.Y., Munoz-Medina,
 745 J.E., Gonzalez-Bonilla, C.R., Arias, C.F., Lewis-Ximenez, L.L., Baylis, S.A., Chieppe, A.O., Aguiar, S.F.,
 746 Fernandes, C.A., Lemos, P.S., Nascimento, B.L.S., Monteiro, H.A.O., Siqueira, I.C., de Queiroz, M.G.,
 747 de Souza, T.R., Bezerra, J.F., Lemos, M.R., Pereira, G.F., Loudal, D., Moura, L.C., Dhalia, R., Franca,
 748 R.F., Magalhaes, T., Marques, E.T., Jr., Jaenisch, T., Wallau, G.L., de Lima, M.C., Nascimento, V., de
 749 Cerqueira, E.M., de Lima, M.M., Mascarenhas, D.L., Neto, J.P.M., Levin, A.S., Tozetto-Mendoza, T.R.,
 750 Fonseca, S.N., Mendes-Correa, M.C., Milagres, F.P., Segurado, A., Holmes, E.C., Rambaut, A., Bedford,
 751 T., Nunes, M.R.T., Sabino, E.C., Alcantara, L.C.J., Loman, N.J., Pybus, O.G., 2017. Establishment and
 752 cryptic transmission of Zika virus in Brazil and the Americas. *Nature* 546, 406-410.

753 Faye, O., Freire, C.C., Iamarino, A., Faye, O., de Oliveira, J.V., Diallo, M., Zannotto, P.M., Sall, A.A., 2014.
 754 Molecular evolution of Zika virus during its emergence in the 20(th) century. *PLoS Negl Trop Dis* 8,
 755 e2636.

756 Gao, D., Lou, Y., He, D., Porco, T.C., Kuang, Y., Chowell, G., Ruan, S., 2016. Prevention and Control of Zika as
 757 a Mosquito-Borne and Sexually Transmitted Disease: A Mathematical Modeling Analysis. *Sci Rep* 6,
 758 28070.

759 Garcia-Luna, S.M., Weger-Lucarelli, J., Ruckert, C., Murrieta, R.A., Young, M.C., Byas, A.D., Fauver, J.R.,
 760 Perera, R., Flores-Suarez, A.E., Ponce-Garcia, G., Rodriguez, A.D., Ebel, G.D., Black, W.C.t., 2018.
 761 Variation in competence for ZIKV transmission by *Aedes aegypti* and *Aedes albopictus* in Mexico.
 762 *PLoS Negl Trop Dis* 12, e0006599.

Giovanetti, M., Milano, T., Alcantara, L.C., Carcangiu, L., Cella, E., Lai, A., Lo Presti, A., Pascarella, S.,
 Zehender, G., Angeletti, S., Ciccozzi, M., 2016. Zika Virus spreading in South America: Evolutionary
 analysis of emerging neutralizing resistant Phe279Ser strains. *Asian Pac J Trop Med* 9, 445-452.
 Griffiths, R.C., Tavaré, S., 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R
 Soc Lond B Biol Sci* 344, 403-410.
 Gubler, D.J., Vasilakis, N., Musso, D., 2017. History and Emergence of Zika Virus. *J Infect Dis* 216, S860-S867.
 Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and
 methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst
 Biol* 59, 307-321.
 Han, J.F., Jiang, T., Ye, Q., Li, X.F., Liu, Z.Y., Qin, C.F., 2016. Homologous recombination of Zika viruses in the
 Americas. *J Infect* 73, 87-88.
 Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., Gibson, T.J., 1998. Multiple sequence alignment
 with Clustal X. *Trends Biochem Sci* 23, 403-405.
 Kass, R.E., Raftery, A.E., 1995. Bayes factors. *Journal of the american statistical association* 90, 773-795.
 Lana, R.M., Carneiro, T.G., Honório, N.A., Codeco, C.T., 2014. Seasonal and nonseasonal dynamics of *Aedes
 aegypti* in Rio de Janeiro, Brazil: fitting mathematical models to trap data. *Acta Trop* 129, 25-32.
 Lazear, H.M., Diamond, M.S., 2016. Zika Virus: New Clinical Syndromes and Its Emergence in the Western
 Hemisphere. *J Virol* 90, 4864-4875.
 Lednicky, J., Beau De Rochars, V.M., El Badry, M., Loeb, J., Telisma, T., Chavannes, S., Anilis, G., Cella, E.,
 Ciccozzi, M., Rashid, M., Okech, B., Salemi, M., Morris, J.G., Jr., 2016. Zika Virus Outbreak in Haiti in
 2014: Molecular and Clinical Data. *PLoS Negl Trop Dis* 10, e0004687.
 Lemey, P., Rambaut, A., Drummond, A.J., Suchard, M.A., 2009. Bayesian phylogeography finds its roots.
PLoS Comput Biol 5, e1000520.
 Lemey, P., Rambaut, A., Welch, J.J., Suchard, M.A., 2010. Phylogeography takes a relaxed random walk in
 continuous space and time. *Molecular biology and evolution* 27, 1877-1885.

788 Liang, D., Leung, R.K.K., Lee, S.S., Kam, K.M., 2017. Insights into intercontinental spread of Zika virus. *PLoS*
789 *One* 12, e0176710.

790 Marchette, N.J., Garcia, R., Rudnick, A., 1969. Isolation of Zika virus from *Aedes aegypti* mosquitoes in
791 Malaysia. *Am J Trop Med Hyg* 18, 411-415.

792 Metsky, H.C., Matranga, C.B., Wohl, S., Schaffner, S.F., Freije, C.A., Winnicki, S.M., West, K., Qu, J., Baniecki,
793 M.L., Gladden-Young, A., Lin, A.E., Tomkins-Tinch, C.H., Ye, S.H., Park, D.J., Luo, C.Y., Barnes, K.G.,
794 Shah, R.R., Chak, B., Barbosa-Lima, G., Delatorre, E., Vieira, Y.R., Paul, L.M., Tan, A.L., Barcellona,
795 C.M., Porcelli, M.C., Vasquez, C., Cannons, A.C., Cone, M.R., Hogan, K.N., Kopp, E.W., Anzinger, J.J.,
796 Garcia, K.F., Parham, L.A., Ramirez, R.M.G., Montoya, M.C.M., Rojas, D.P., Brown, C.M., Hennigan, S.,
797 Sabina, B., Scotland, S., Gangavarapu, K., Grubaugh, N.D., Oliveira, G., Robles-Sikisaka, R., Rambaut,
798 A., Gehrke, L., Smole, S., Halloran, M.E., Villar, L., Mattar, S., Lorenzana, I., Cerbino-Neto, J., Valim, C.,
799 Degraeve, W., Bozza, P.T., Gnirke, A., Andersen, K.G., Isern, S., Michael, S.F., Bozza, F.A., Souza, T.M.L.,
800 Bosch, I., Yozwiak, N.L., MacInnis, B.L., Sabeti, P.C., 2017. Zika virus evolution and spread in the
801 Americas. *Nature* 546, 411-415.

802 Mittal, R., Nguyen, D., Debs, L.H., Patel, A.P., Liu, G., Jhaveri, V.M., Si, S.K., Mittal, J., Bandstra, E.S., Younis,
803 R.T., Chapagain, P., Jayaweera, D.T., Liu, X.Z., 2017. Zika Virus: An Emerging Global Health Threat.
804 *Front Cell Infect Microbiol* 7, 486.

805 Moreira-Soto, A., Carneiro, I.O., Fischer, C., Feldmann, M., Kummerer, B.M., Silva, N.S., Santos, U.G., Souza,
806 B., Liborio, F.A., Valenca-Montenegro, M.M., Laroque, P.O., da Fontoura, F.R., Oliveira, A.V.D.,
807 Drosten, C., de Lamballerie, X., Franke, C.R., Drexler, J.F., 2018. Limited Evidence for Infection of
808 Urban and Peri-urban Nonhuman Primates with Zika and Chikungunya Viruses in Brazil. *mSphere* 3,
809 00523-00517.

810 Musso, D., Bossin, H., Mallet, H.P., Besnard, M., Broult, J., Baudouin, L., Levi, J.E., Sabino, E.C., Ghawche, F.,
811 Lanteri, M.C., Baud, D., 2018. Zika virus in French Polynesia 2013-14: anatomy of a completed
812 outbreak. *Lancet Infect Dis* 18, e172-e182.

813 Olawoyin, O., Kribs, C., 2018. Effects of multiple transmission pathways on Zika dynamics. *Infect Dis Model*
814 3, 331-344.

815 Passos, S.R.L., Borges Dos Santos, M.A., Cerbino-Neto, J., Buonora, S.N., Souza, T.M.L., de Oliveira, R.V.C.,
816 Vizzoni, A., Barbosa-Lima, G., Vieira, Y.R., Silva de Lima, M., Hokerberg, Y.H.M., 2017. Detection of
817 Zika Virus in April 2013 Patient Samples, Rio de Janeiro, Brazil. *Emerg Infect Dis* 23, 2120-2121.

818 Pettersson, J.H., Bohlin, J., Dupont-Rouzeyrol, M., Brynildsrud, O.B., Alfsnes, K., Cao-Lormeau, V.M., Gaunt,
819 M.W., Falconar, A.K., de Lamballerie, X., Eldholm, V., Musso, D., Gould, E.A., 2018. Re-visiting the
820 evolution, dispersal and epidemiology of Zika virus in Asia. *Emerg Microbes Infect* 7, 79.

821 Posada, D., 2008. jModelTest: phylogenetic model averaging. *Molecular biology and evolution* 25, 1253-
822 1256.

823 Rambaut, A., Lam, T.T., Max Carvalho, L., Pybus, O.G., 2016. Exploring the temporal structure of
824 heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* 2, vew007.

825 Sebastian, U.U., Ricardo, A.V.A., Alvarez, B.C., Cubides, A., Luna, A.F., Arroyo-Parejo, M., Acuna, C.E.,
826 Quintero, A.V., Villareal, O.C., Pinillos, O.S., Vieda, E., Bello, M., Pena, S., Duenas-Castell, C.,
827 Rodriguez, G.M.V., Ranero, J.L.M., Lopez, R.L.M., Olaya, S.G., Vergara, J.C., Tandazo, A., Ospina, J.P.S.,
828 Leyton Soto, I.M., Fowler, R.A., Marshall, J.C., 2017. Zika virus-induced neurological critical illness in
829 Latin America: Severe Guillain-Barre Syndrome and encephalitis. *J Crit Care* 42, 275-281.

830 Simon-Loriere, E., Holmes, E.C., 2011. Why do RNA viruses recombine? *Nat Rev Microbiol* 9, 617-626.

831 Simonin, Y., van Riel, D., Van de Perre, P., Rockx, B., Salinas, S., 2017. Differential virulence between Asian
832 and African lineages of Zika virus. *PLoS Negl Trop Dis* 11, e0005821.

833 Strimmer, K., von Haeseler, A., 1997. Likelihood-mapping: a simple method to visualize phylogenetic
834 content of a sequence alignment. *Proc Natl Acad Sci U S A* 94, 6815-6819.

835 Towers, S., Brauer, F., Castillo-Chavez, C., Falconar, A.K.I., Mubayi, A., Romero-Vivas, C.M.E., 2016. Estimate
836 of the reproduction number of the 2015 Zika virus outbreak in Barranquilla, Colombia, and
837 estimation of the relative role of sexual transmission. *Epidemics* 17, 50-55.

838 Weaver, S.C., Costa, F., Garcia-Blanco, M.A., Ko, A.I., Ribeiro, G.S., Saade, G., Shi, P.Y., Vasilakis, N., 2016.
839 Zika virus: History, emergence, biology, and prospects for control. *Antiviral Res* 130, 69-80.

840 Zanluca, C., Melo, V.C., Mosimann, A.L., Santos, G.I., Santos, C.N., Luz, K., 2015. First report of
841 autochthonous transmission of Zika virus in Brazil. *Mem Inst Oswaldo Cruz* 110, 569-572.

842 Zehender, G., Veo, C., Ebranati, E., Carta, V., Roviola, F., Percivalle, E., Moreno, A., Lelli, D., Calzolari, M.,
843 Lavazza, A., Chiapponi, C., Baioni, L., Capelli, G., Ravagnan, S., Da Rold, G., Lavezzo, E., Palu, G.,
844 Baldanti, F., Barzon, L., Galli, M., 2017. Reconstructing the recent West Nile virus lineage 2 epidemic
845 in Europe and Italy using discrete and continuous phylogeography. *PLoS One* 12, e0179679.

846 Zhang, Q., Sun, K., Chinazzi, M., Pastore, Y.P.A., Dean, N.E., Rojas, D.P., Merler, S., Mistry, D., Poletti, P.,
847 Rossi, L., Bray, M., Halloran, M.E., Longini, I.M., Jr., Vespignani, A., 2017. Spread of Zika virus in the
848 Americas. *Proc Natl Acad Sci U S A* 114, E4334-E4343.

849

850